MODELING BIOLOGICAL PROCESSES IN GENOME-WIDE ASSOCIATION STUDIES USING REGULARIZED REGRESSION

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Gabriel Hoffman

May 2013

MODELING BIOLOGICAL PROCESSES IN GENOME-WIDE ASSOCIATION STUDIES USING REGULARIZED REGRESSION

Gabriel Hoffman, Ph.D.

Cornell University 2013

Genome-wide association studies (GWAS) have become a a widely adopted approach to identify genetic variation that produces variation in complex phenotype. Standard statistical methods are able to identify strong associations in these datasets, but more sophisticated statistical methods that model complex aspects of the biological data can identify weaker associations and further elucidate the underlying molecular biology. We develop and apply statistical methods that explicitly model two aspects of GWAS data using two complementary forms of regularized regression. First, we model the polygenic architecture of complex phenotypes using feature selection methods in a penalized regression framework. We propose novel algorithmic, computational and heuristic approaches in order to produce a method that scales to high dimensional GWAS data and increases power to detect weak associations that are not detectable by standard tests. Second, we model the covariance between individuals due to kinship and population structure using a linear mixed model that regularizes the statistical contribution of a metric of ancestry. Linear mixed models have been widely adopted for analysis of GWAS data, but their theoretical properties have not been examined in this context. We formalize the statistical properties of the linear mixed model, develop a novel interpretation in relation to population genetics, and propose a novel low rank linear mixed model that learns the dimensionality of the correction for kinship and population structure

from the data. Finally, we combine these two complementary regularized regression models into a penalized linear mixed model. We develop a unified model incorporating a novel algorithm with novel approaches to tuning non-convex penalties and determining the optimal stopping point in the regularization path. Leveraging recent work on assessing significance of selected features, we produce a well-principled and scalable statistical method applicable to feature selection, hypothesis testing and prediction in many contexts.

BIOGRAPHICAL SKETCH

Gabriel grew up in Elkins Park, Pennsylvania where he graduated from Cheltenham High School in 2003. He attended the University of Maryland, College Park where he graduated *cum laude* with Honors with a B.S. in Cell, Molecular Biology and Genetics and a minor in Computer Science. While at the University of Maryland, Gabriel worked under Dr. Charles Delwiche studying the molecular evolution of eukaryotic algae. In 2007, Gabriel began graduate studies at Cornell University under Dr. Jason Mezey where he developed and applied statistical methods to identify the genetic variation underlying complex traits.

To my family

ACKNOWLEDGEMENTS

I would first like to thank my family for their constant support. I wish to acknowledge members of the Mezey lab, especially Benjamin Logsdon, Larsson Omberg and Anthony Greenberg, for helping to motivate much of the work presented here. I would like to thank Andrew Clark, Adam Siepel, Martin Wells and Jim Booth for guidance and helpful discussions. I would also like to thank my funding sources from Cornell including the Presidential Life Sciences Fellowship, Genetics and Development Training Grant, the Cornell Center for Comparative and Population Genetics Fellowship and NSF grant DEB0922432. Finally, I have to thank my advisor, Jason Mezey, for his patience and encouragement.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 Genome-Wide Association Studies

Understanding how genetic variation leads to variation in phenotype has long been the goal of quantitative genetics. Establishing a link between genotype and phenotype elucidates the molecular mechanisms underlying the phenotype and is a broad interest in medical, agricultural and model-organism genetics. Genotyping technologies assaying over 1 million common single nucleotide polymorphisms (SNPs) in humans has facilitated hundreds of genome-wide association studies (GWAS) [68]. These studies have advanced the understanding the molecular biology of common diseases [80, 88, 140, 164, 178, 193] as well as drug response [229], metabolic [38, 179, 180, 185] and anthropometric phenotypes [95, 174]. This research has laid the foundation for new approaches to treatment by identifying genetic variants associated with drug response [33, 76, 78, 135, 183, 228], characterizing current and potential drug targets [143, 194], and exploring the application of existing drugs to other diseases [163].

1.1.1 Statistical methods model different aspects of biology

Standard analyses of GWAS datasets perform a separate test of association for each genetic marker while including known confounding variables such as sex. Single marker tests essentially consider the correlation between a single genetic marker and a phenotype of interest. Such tests have been widely applied since

they are able to identify strong associations, they are scalable to very large datasets, and they use a simple and interpretable statistical model. Yet these methods do not model complexities of biological datasets and more sophisticated methods have the potential to increase statistical power and decrease the false discovery rate. To this end, much recent attention has focused on biological complexities such as epistasis [21, 104, 119, 127, 133, 149, 159, 167, 191, 214, 223], gene $\times$ environment interaction [6, 58, 116, 161, 186, 231], the effect of rare variants [7, 98, 102, 120, 134, 151, 208, 220], gene- or pathway-based disease models [21, 27, 73, 101, 105, 111], the polygenic architecture of complex traits [9, 21, 65, 72, 114, 133, 227], and the confounding effect of kinship and population structure [83, 84, 108, 109, 146, 150, 182, 230].

While there has been much work on each of these aspects of GWAS data, the complexity and ever-increasing size of relevant datasets necessitates sophisticated models grounded in biology that are still computationally scalable. The complexities of the underlying biological system can be modeled by a range of disparate approaches. Yet regularized regression stands out as being applicable to each aspect discussed above [9, 27, 84, 208, 214, 231] and further being able to incorporate prior biological knowledge [21, 118, 159]. Here, we apply two complementary forms of regularized regression to modeling the polygenic architecture of complex traits (Chapter 2) and correcting for kinship and population structure (Chapter 3). Finally, we combine these methods into a unified statistical method that models both of these aspects of GWAS data (Chapter 4).

1.2 Regularized regression

In general, *regularization* introduces into a statistical model some prior knowledge of the system underlying the observed data in order to avoid overfitting, make the model computationally tractable, or explicitly conform to the prior. In regularized regression, this takes the form of a prior on regression coefficients. In the maximum likelihood context considered here, two complementary approaches to regularized regression take the form of a penalized regression and a random effects model.

1.2.1 Penalized regression

In penalized regression, a prior is placed on the regression coefficients and the corresponding penalized likelihood is maximized with respect to these coefficients, conditional on a fixed tuning parameter value. An illustrative example is the so-called ridge regression [64, 70] of the form

$$L_{\text{Ridge}}(\boldsymbol{\beta}|\mathbf{y}) \;=\; \mathcal{N}(\mathbf{y}|\mathbf{X}\boldsymbol{\beta}, \sigma_e^2)\, \mathcal{N}(\boldsymbol{\beta}|0, \lambda) \tag{1.1}$$

in which both the data likelihood and prior are Gaussian. In general, the data likelihood and prior can be a chosen from number of distributions so that the penalized log-likelihood can be expressed as a data log-likelihood term, and a penalty (i.e. log prior) term

$$\ell_{penalized}(\boldsymbol{\beta}|\mathbf{y}) \;=\; \ell(\boldsymbol{\beta}|\mathbf{y}) - p_\lambda(\boldsymbol{\beta}) \tag{1.2}$$

indexed by a tuning parameter λ. There is an extensive literature concerning how the choice of penalty affects the statistical properties of $\hat{\boldsymbol{\beta}}$ [5, 20, 41, 42, 67, 187].

Table 1.1: Properties of feature selection penalties

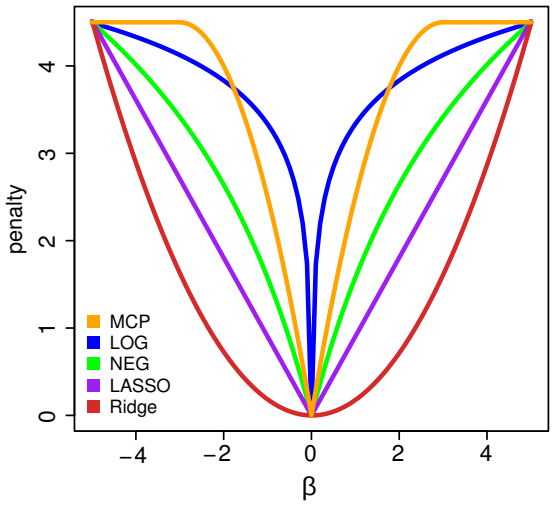
Penalty	Feature selection	Convex	Oracle	Tuning parameters	Reference
Ridge	no	yes	no	1	[70]
LASSO	yes	yes	no	1	[187]
MCP	yes	no	yes	2	[221]
LOG	yes	no	yes	2	[124]
NEG	yes	no	yes	2	[54]

Feature selection in penalized regression

In many applications of statistics the goal of an analysis is often to identify a subset of features that are relevant to the response while discarding all other features. Statistically, this is achieved by estimating the regression coefficients of relevant features while setting the coefficients of irrelevant features to exactly zero. This objective can be formulated in the context of penalized regression by using a penalty that satisfies certain conditions and setting the tuning parameter so that only a subset of the coefficients have nonzero coefficients. In order to perform feature selection, the penalty function must be non-negative, non-decreasing, and non-differentiable at the origin [42] (Figure 1.1, Table 1.1) . The most widely used of such penalties is the LASSO ('least absolute shrinkage and selection operator') [187]. The fact the LASSO is convex means that it selects a unique set of features with nonzero coefficients for a given value of the tuning parameter when there are more samples (n) than features (p) [187, 189], and selects a set of features that span a unique subspace when $p > n$ [189].

Yet the convexity of the LASSO comes at a cost. For high-dimensional datasets,

Figure 1.1: Penalty functions used in penalized regression



the LASSO is not guaranteed to recover the sparse structure of the true model even asymptotically in the sample size [226]. However, other penalties satisfy the so-called 'oracle property' whereby the coefficient estimates have the same mean and covariance asymptotically as if the relevant features were known beforehand [42]. Such penalties have a derivative that approaches zero as the coefficient approaches infinity [42] and include MCP, LOG and NEG. Thus these penalties have desirable statistical properties but practical and computational issues arise due to the fact that they are are nonconvex and have two tuning parameters. We address these issues to produce a scalable framework for nonconvex penalties that can outperform existing methods (Chapters 2 and 4).

1.2.2 Random effects models

In a random (or mixed) effects model, a prior is placed on the regression coefficients and the likelihood is maximized with respect to the variance components,

while the coefficients are integrated out. An illustrative example is the standard linear mixed model [145] of the form

$$L_{\text{LMM}}(\sigma_a^2, \sigma_e^2|\mathbf{y}) \;\; = \;\; \int \mathcal{N}(\mathbf{y}|\mathbf{X}\boldsymbol{\beta}, \sigma_e^2)\; \mathcal{N}(\boldsymbol{\beta}|0, \sigma_a^2)\; d\boldsymbol{\beta} \tag{1.3}$$

where both the data likelihood and prior are Gaussian. This model is markedly different from ridge regression due to the integral and the now essential role of the variance components. Other forms of random effects models are possible either by using a complementary data likelihood and prior distribution [100], using numerical integration [50], or approximating the integral [77, 196, 206].

Widely used in applied statistics, random effects models allow fitting of overdetermined systems, shrink coefficient estimates toward the prior mean and reduce the variance of estimates of fixed effects [145, 206]. In machine learning, these models are termed Gaussian process regression (GPR) [157] and are expressed as

$$L_{\text{GPR}}(\sigma_a^2, \sigma_e^2|\mathbf{y}) \;\; = \;\; \int \mathcal{N}(\mathbf{y}|\boldsymbol{\alpha}, \sigma_e^2)\; \mathcal{N}(\boldsymbol{\alpha}|0, \mathbf{K}\sigma_a^2)\; d\boldsymbol{\alpha} \tag{1.4}$$

which is equivalent to (1.3) when $\mathbf{K} = \mathbf{X}\mathbf{X}^T$. GPR is thus a 'kernel method' where the kernel matrix, $\mathbf{K}$, represents a metric of similarity between all pairs of samples [157].

The linear mixed model has recently been widely adopted in statistical genetics in order to account for the covariance between samples [83, 84, 108, 109, 146, 152, 182, 230] and many similarity metrics have been proposed [8, 83, 96, 97, 144, 150, 165, 173, 215]. We formalize and interpret the relationship between (1.3) and (1.4) in the context of statistical genetics and develop a novel method that extends this framework (Chapters 3 and 4).

CHAPTER 2

PUMA: A UNIFIED FRAMEWORK FOR PENALIZED MULTIPLE REGRESSION ANALYSIS OF GWAS DATA

2.1 Abstract

Penalized Multiple Regression (PMR) can be used to discover novel disease associations in GWAS datasets. In practice, proposed PMR methods have not been able to identify well-supported associations in GWAS that are undetectable by standard association tests and thus these methods are not widely applied. Here, we present a combined algorithmic and heuristic framework for PUMA (Penalized Unified Multiple-locus Association) analysis that solves the problems of previously proposed methods including computational speed, poor performance on genome-scale simulated data, and identification of too many associations for real data to be biologically plausible. The framework includes a new minorize-maximization (MM) algorithm for generalized linear models (GLM) combined with heuristic model selection and testing methods for identification of robust associations. The PUMA framework implements the penalized maximum likelihood penalties previously proposed for GWAS analysis (i.e. Lasso, Adaptive Lasso, NEG, MCP), as well as a penalty that has not been previously applied to GWAS (i.e. LOG). Using simulations that closely mirror real GWAS data, we show that our framework has high performance and reliably increases power to detect weak associations, while existing PMR methods can perform worse than single marker testing in overall performance. To demonstrate the empirical value of PUMA, we analyzed GWAS data for type 1 diabetes, Crohns's disease, and rheumatoid arthritis, three autoimmune diseases

from the original Wellcome Trust Case Control Consortium. Our analysis replicates known associations for these diseases and we discover novel etiologically relevant susceptibility loci that are invisible to standard single marker tests, including 6 novel associations implicating genes involved in pancreatic function, insulin pathways and immune-cell function in type 1 diabetes; 3 novel associations implicating genes in pro- and anti-inflammatory pathways in Crohn's disease; and 1 novel association implicating a gene involved in apoptosis pathways in rheumatoid arthritis. We provide software for applying our PUMA analysis framework.

2.2 Introduction

Genome-wide association studies (GWAS) have identified many susceptibility loci underlying the molecular etiology of complex diseases [68]. These studies have been responsible for the discovery of many individual genes that contribute to disease risk [3, 10, 38, 46, 49, 95, 176, 195, 205], for discoveries on the front line of personalized medicine [76, 228], and for discovering novel pathways important for the progression of complex heritable diseases [201]. The expense of each GWAS that is capable of finding well-supported disease loci is considerable and, as a consequence, each robust and interpretable association discovered in a GWAS is valuable, not only from the point of view of scientific discovery but also in terms of return on investment [4, 175]. A clear picture that has an important bearing on the investment-discovery tradeoff in GWAS experiments is that the associations identified to date generally explain only a small to moderate fraction of total heritability [121, 122]. Recent analyses have suggested that a considerable amount of this 'missing' heritability can be accounted for by

rare variants or variants with weak effects [141, 215, 217]. This suggests that there is an opportunity to identify more risk loci through studies that require even greater investment, by including larger sample sizes and/or by incorporating higher genetic marker coverage of the genome by using next-generation sequencing (NGS). The novel associations discovered by large consortia GWAS studies support this supposition [3, 38, 49, 95]. Another complementary strategy that leverages both the current and future investment in GWAS experiments is the application of new statistical analyses that can reliably identify weaker associations [72, 91, 127, 160, 200]. Although there has been an explosion of methods in this area [21, 133], few have produced robustly supported associations that are not detectable by single marker tests of association [21, 68, 133, 178, 193].

Here we report a general framework for applying a family of GWAS analysis methods that is extremely promising for detection of weak associations yet has not been widely applied to learn novel biology from GWAS datasets: penalized multiple regression (PMR) methods. PMR methods work by simultaneously incorporating tens to hundreds of thousands of genetic markers in a single statistical model where a penalty is incorporated to force most marker regression coefficients to be exactly zero, so that only a small fraction are estimated to make a contribution to disease risk [9, 39, 57, 65, 72, 103, 114, 192, 209, 214, 227]. By jointly analyzing markers, PMR methods are able to consider the correlation of each marker with the phenotype, conditional on the effect of all other markers. This can increase the power to detect weak associations compared to single marker methods due to the smaller residual variance and the fact that the conditional correlation of a marker with the phenotype (i.e. the correlation of a given marker with the phenotype once the estimated effect of all other relevant mark-

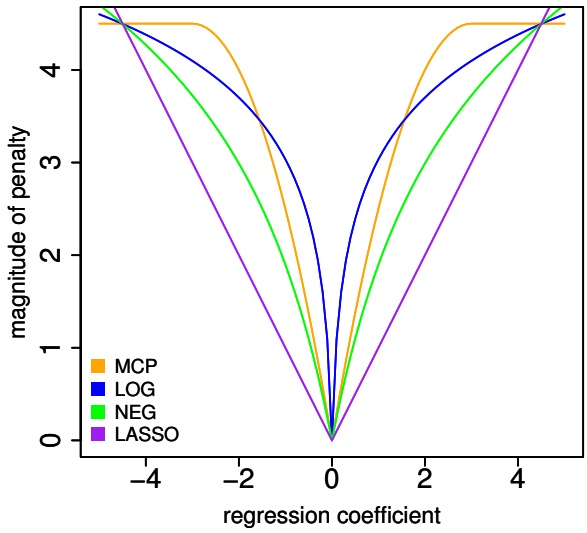
ers has been subtracted out) can often be substantially higher than the marginal correlation [64]. The latter effect is a consequence of non-zero correlation structure between associated markers when the underlying genetic architecture is polygenic [22]. These methods therefore model the underlying biology more accurately than single marker tests, by explicitly modeling the polygenic architecture of complex phenotypes to account for the effects of multiple susceptibility loci. They also leverage the same type of statistical model used in single marker testing methods that have demonstrated reliability in the identification of strong associations [68, 178, 193]. Yet, despite theoretical power of PMR methods, the large body of statistical literature exploring their theoretical properties (see reviews [20, 44]), and the recent interest in the methods development community [9, 39, 57, 65, 72, 103, 114, 192, 209, 214, 227], these methods have not been successful in GWAS analysis. This is due to a combination of limitations: 1) inability to scale for very large GWAS datasets [9, 57, 72, 103], 2) poor performance on simulated data [72, 209], 3) they often find too many 'hits' to be biologically plausible for a given GWAS sample size [72], and 4) they do not identify novel, well-supported associations that are not detectable by standard methods [72, 209].

In order to address these issues, we present a combined algorithmic and heuristic framework for PUMA (Penalized Unified Multiple-locus Association) analysis that optimizes these methods for reliable detection of weak associations when applied to large GWAS datasets. The complete PUMA framework includes an extremely efficient implementation of a new minorize-maximization (MM) algorithm [75] for generalized linear models (GLM) [125], a theoretically motivated data-adaptive heuristic approach to determine penalty strength and

for model selection, and *post hoc* methods for assessing the rank of identified associations. Within PUMA, we implement all sparse feature selection, penalized regression approaches proposed for GWAS analysis to date, including four penalties implemented in a maximum likelihood framework (i.e. Lasso, Adaptive Lasso, NEG, MCP), as well as theoretically justified penalties that have not been previously applied to GWAS (i.e. LOG) (Figure 2.1). We demonstrate the power of our framework for detecting weaker associations that are invisible to individual marker testing through analysis of simulated GWAS data that mirror observations from analyses of real GWAS data. We also demonstrate that our approaches correct issues with all current PMR methods where software is available for GWAS analysis, where we find that all of these currently available PMR GWAS methods can perform worse than single marker testing for our simulation conditions. As an illustration of the value of PUMA for mining existing GWAS data for novel associations, we apply these methods to the original Wellcome Trust Case Control Consortium (WTCCC) [205] GWAS datasets for type 1 diabetes, Crohn's disease and rheumatoid arthritis. Our re-analysis identifies weak associations that implicate additional susceptibility loci for these autoimmune diseases, which did not appear significant by standard single marker tests of association in these datasets, yet were 1) identified in an independent GWAS of the same phenotype that did not include WTCCC data, 2) previously known to play a role in disease etiology, or 3) known to function in a relevant biological pathway. Our results demonstrate that appropriately tuned PMR methods can provide a complementary approach to large meta-analyses [3, 10, 38, 46, 49, 95, 176] to identify susceptibility loci with weak associations. We also provide a discussion concerning how the framework can be extended to perform penalized analysis of epistasis, to incorporate mixed model analy-

sis, and to address challenges of genome-wide genotypes provided by whole-genome next-generation sequencing.

Figure 2.1: Penalty functions on the magnitude of the regression coefficients implemented in the PUMA framework. A parameter determines the slope near the origin for all penalties, while MCP, LOG and NEG have an additional tuning parameter determining the rate at which the derivative of the penalty tails off to zero.



2.3 Results

2.3.1 PUMA is a scalable framework for GWAS analysis

The methods implemented in our PUMA framework are orders of magnitude faster than existing software when assigned identical computational tasks and no pre-screening of markers is performed (Table 2.2). This substantial boost in computational speed allows PUMA to perform a dense two-dimensional search of tuning parameter values for non-convex penalties (i.e. MCP, NEG, LOG) and

examine upwards of 1 million total modes of the likelihood surface for simulated case/control dataset of 5,000 individuals and 650K genetic markers in less than 24 hours on a 6 core Intel® Xeon® W3690 @ 3.47GHz with 12 Gb memory when a pre-screening p-value cutoff of 0.01 from single marker analysis is applied (Table 2.1). This is a huge improvement compared to existing software for non-convex PMR methods [9, 72] which only examine a single mode.

2.3.2 Theoretical and empirical justification for pre-screening markers

While pre-screening markers based on a p-value cutoff may initially seem to detract from the purpose of a multiple-locus analysis, it is supported by statistical theory, is necessary for large scale analysis and has almost no impact on the set of markers identified as associated. In a seminal paper, Fan and Lv [43] demonstrate that pre-screening by ranking the marginal correlation of each variable with the response will retain the relevant variable asymptotically with probability tending to 1. Fan and Song [44] extend this result to generalized linear models. Moreover, Tibshirani, et al. [188] and El Ghaoui, et al. [52] establish exact pre-screening methods for linear and logistic Lasso models where relevant variables are guaranteed to be retained for finite sample sizes and demonstrate that the number of variables can be reduced by up to 3 orders of magnitude. Intuitively, both the asymptotic [43, 44] and exact pre-screening methods [52, 188] rely on the fact that a variable is unlikely to have a very small marginal correlation with the response but a large and very significant conditional correlation for a particular sample size when the relevant variables explain only a small

Table 2.1: Run times for PUMA methods and other available software. For a typical simulated data set with 5000 individuals, 650K markers and a pre-screening p-value threshold of 0.01, we report the run times, and the number of total and unique models examined by our methods (top) and available methods using standard / default settings (bottom). We list the number of models assessed during a single run of a method where a model is defined by the set of markers with distinct nonzero coefficients and the number of unique models counts the number of sets of distinct markers, where we note that the metrics reported can vary substantially between datasets. Lasso and Adaptive Lasso are convex and have a single tuning parameter, so relatively few models are examined during the search. For convex penalties, each distinct tuning parameter value produces a model, although another tuning parameter value can cause the coefficients to change but still produce the same set of markers with nonzero coefficients. Thus the number of models examined is larger than the number of unique models. MCP, LOG and NEG penalties are non-convex and have two tuning parameters and were applied with 100 marker reorderings, so they produce orders of magnitude more total and unique models. We note that 1D-MCP is faster than 2D-MCP as the former fixes the value of one tuning parameter. We note that HyperLasso [72] can be extremely computationally expensive for large datasets, so that the time we report is based on analysis of the pre-screened dataset where pre-screening step must be implemented separately. Ayers and Cordell [9] do not provide software but proposes an approach using the *grpreg* package in R.

Method	Run time	# of models	# of unique models
Lasso	33 s	156	59
Adaptive Lasso	5 s	21	13
LOG	6 hrs	$\sim 700,000$	$\sim 4,000$
NEG	5 hrs	$\sim 500,000$	$\sim 10,000$
1D-MCP	21 min	$\sim 800,000$	21
2D-MCP	14 hrs	$\sim 1,000,000$	$\sim 5,000$
Mendel [209]	66 s	1	1
HyperLasso [72]	1 hr	1	1
perm-MCP [9]	1 hr	1	1

fraction of the variation in the response. Moreover, pre-screening is often computationally necessary because storing 650K markers for 5000 samples requires

Table 2.2: Run times for PUMA and other available software for identical analyses. For a typical simulated data set with 650K markers, no pre-screening of markers and sample sizes, n, of 1000, 2000 and 5000, we report run times for available software and PUMA performing the same analyses. For Lasso, we had Mendel and PUMA perform a search of tuning parameter space in order to return $1.5\sqrt{n}$ markers with nonzero coefficients. For NEG, we set HyperLasso to its default tuning parameter values and ran PUMA with the same values. For MCP, we set grpreg and PUMA to perform a search of tuning parameter space in order to return $1.5\sqrt{n}$ markers with nonzero coefficients, where γ was set to 30 as per Ayers and Cordell [9]. Analysis was performed on an 8 core Intel® Xeon® E5520 @ 2.27GHz with 32 Gb memory. NA indicates the program crashed due to insufficient memory; we note that this is due to technical limitations of Mendel and R, in which grpreg runs.

	Sample size		
Method	1000	2000	5000
Lasso	2m 11s	5m 55s	14m 45s
NEG	1.2s	2.2s	9.8s
MCP	4.7s	8.2s	29.2s
Mendel [209] (Lasso)	9m 50s	NA	NA
HyperLasso [72] (NEG)	52m 24s	4h 16m	20h 3m
grpreg [9] (MCP)	3h 52m	NA	NA

26 Gb of memory. Finally, we note that pre-screening is used by previous applications of PMR methods to GWAS data [72, 209] in order to handle genome-scale data.

We use a pre-screening p-value cutoff based on single marker analysis, because 1) it retains all relevant variables asymptotically [43, 44], 2) it approximates the exact methods proposed for Lasso [52, 188], which cannot be easily adapted to other penalties, 3) it reduces memory requirements so that very large datasets can be analyzed on a high-end desktop computer, 4) it substantially reduces the computational burden, 5) by using a p-value it is naturally calibrated to the

sample size and the fraction of variation in the response being explained, and 6) it has very little empirical effect on the results.

We demonstrate this final and most important point in two complementary simulation studies. First we consider a simple two-step forward regression method, which is known to approximate penalized multiple regression [37, 63] and, under a range of biologically motivated simulation conditions, demonstrate that variables that do not cross an initial p-value threshold have a very low probability of being significant in the second step (Figure A.1). Second we demonstrate that the pre-screening has no noticeable effect on the performance of Lasso and MCP methods but substantially reduces the computational time (Figure A.2).

2.3.3 Simulated data assessment of the PUMA framework

We analyzed 960 simulated GWAS datasets to assess the performance of our PUMA framework compared to other published methods for PMR GWAS analysis. We note that these simulations, while far more extensive than other published works on PMR GWAS analysis [9, 39, 57, 65, 72, 103, 114, 209, 214, 227] are not meant to be exhaustive or to capture all the possible complexities in a GWAS but rather to: 1) serve as a baseline for comparing GWAS analysis methods and 2) provide an estimate of the expected performance for these methods when applied to GWAS data under relatively ideal experimental conditions. Our goal therefore was not to attempt to model a broad spectrum of possible GWAS data complexities (e.g. stratified experimental sampling schemes, known or cryptic population structure effects on phenotype, relatedness among individuals, measured or latent covariates, etc.) but rather to simulate data that captured

the most basic components of a GWAS experiment (see Methods for details). In simulated data a causal variant is defined as a variant whose coefficient value is nonzero, so that number of minor alleles at this marker contributes to the phenotype. In order to mimic the fact that true causal variants are not available from array-based genotyping, the simulated causal variants were removed from the dataset so that they are not considered by the tests of association. Therefore, just like in all array-based genotyping datasets, our simulations identify associations based on markers in linkage-disequilibrium with the (omitted) causal variant.

2.3.4 Assessment of available software for PMR GWAS

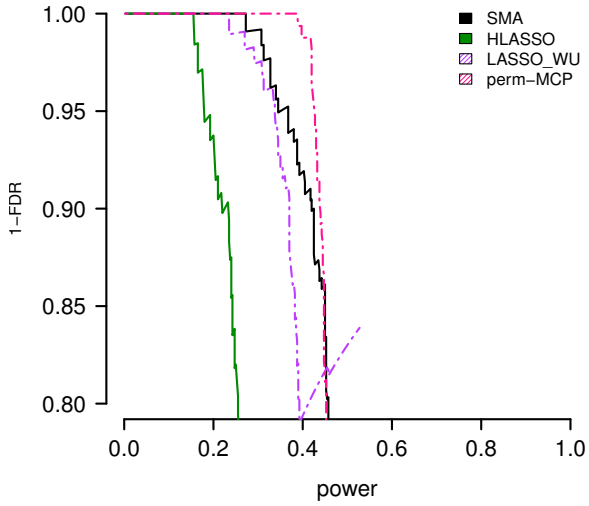
We assessed the performance of PMR methods for which there is available software. We compared the performance of the Lasso penalty from Wu, et al. [209], the NEG penalty as implemented in the HyperLasso program [72], and a permutation-based approach to selecting tuning parameter values for the MCP penalty [9, 221] that we term perm-MCP. We note that we only considered PMR approaches that are designed to handle the specific challenges of GWAS data and that also perform feature selection, such that we do not consider ridge, elastic net, or group-penalties since they set many correlated markers to have nonzero coefficients and thus complicate the generation of interpretable p-values [28, 227]. We also did not consider Markov Chain Monte Carlo (MCMC) approaches [57, 103] since they could not efficiently scale to genome-wide data while exploring a range of tuning parameter values. We ran the HyperLasso program [72] with standard settings (see Appendix A.2). We applied the method of Wu, et al. [209], setting the number of selected markers to the true

number of causal markers in each simulation since Wu, et al. [209] do not specify a criterion for selecting the model size. As a benchmark, we also ran a single marker analysis implemented by applying a logistic regression model to each marker individually. We used a pre-screening p-value cutoff of 0.01 from single marker analysis for the PMR methods to make them computationally tractable.

Simulations indicate that HyperLasso [72] and the Lasso of Wu, et al. [209] are generally less powerful than a standard single marker test (Figure 2.2, A.3-A.10). While Lasso is sometimes comparable or slightly more powerful than a single marker test for low FDR, the performance of the method benefits from the fact that the number of selected markers is set using information not available in real data. Setting the marker number to 10 (the default in the implementation of Wu, et al. [209]) or another arbitrary value results in poor performance and is not competitive with a single marker test (results not shown). The performance of HyperLasso is especially poor as is it suffers from the fact that the choice of tuning parameters has a huge effect on performance, but the method does not implement a search over tuning parameter values. Moreover, HyperLasso does not include a way to evaluate the significance of a selected marker, so we used their default approach of using coefficient values from selected markers to assess performance. Alternatively, perm-MCP was the most powerful in our simulations.

We note that for perm-MCP, by setting the expected false positive rate (eFPR) and using permutations to obtain the value of the tuning parameter based on this rate, perm-MCP generates a single model with relatively few nonzero coefficients while explicitly addressing the multiple testing problem. Yet in practice

Figure 2.2: Simulation results for existing methods. Shown here are representative examples of simulation results for available software including the HyperLasso program [72] (HLASSO), Lasso using the method of Wu, et al. [209] (LASSO_WU) and perm-MCP method [9]. Power is compared to a standard single marker analysis (SMA). Results are shown for 20 replicate datasets from simulations with 5000 individuals, 20 causal markers affecting disease risk and a heritability of 50%. Note that perm-MCP selected very few markers per simulation so the false discovery rate did not exceed 10%.



this result indicates that perm-MCP may assign p-values to only a handful of markers so that the method may not identify any novel associations for a particular dataset. Since the number of nonzero coefficients is directly related to the specified eFPR and the pre-screening cutoff, we examined multiple eFPR values (1×10^{-3},1×10^{-4},1×10^{-5},1×10^{-6},1×10^{-7}) and cutoff values (0.1, 0.01, 0.001), and selected the values the yielded the highest power (eFPR=1×10^{-3}, cutoff=0.001) to present in Figure 2, where other cutoff combinations produce poorer performance (see Figure A.11 for a representative plot showing the results for all cutoffs). We note that the eFPR value is based on the number of markers that pass the pre-screening cutoff, not the total number of markers. Therefore the performance of perm-MCP is sensitive to the eFPR and cutoff val-

ues, yet there is no clear method to optimally specify this value *a priori*. Furthermore, determining the appropriate cutoff for a desired eFPR for correlated high-dimensional data is the subject of current research [53], and its application to permutation methods for selecting a tuning parameter remains an open question. We also note that the performance achieved with PUMA methods does not require the optimal determination of eFPR and pre-screening cutoffs.

In addition, we note that while Ayers and Cordell [9] have previously shown that penalized regression methods can perform well on simulated data, the datasets we address here are orders of magnitude larger. Ayers and Cordell [9] conducted two simulation studies, one with 4000 markers and the other with no more than 228. By considering such a small set of markers, which is not the product of a pre-screening step, they were able to use standard R packages and apply a permutation method to select tuning parameters on the full dataset. Moreover, the multiple testing problem is less severe in their analysis. For the HyperLasso program, Ayers and Cordell [9] selected the tuning parameter as described by Hoggart, et al. [72]. However, using these settings for the genome-scale datasets examined here caused the HyperLasso program to crash (Appendix A.2) and so we use the default program settings. We note that the program worked as expected for smaller datasets. It is unclear whether this problem is an issue with the underlying algorithm or the specifics of the implementation. Thus the difference between the performance of methods in Ayers and Cordell [9] and the current study is the scale of the data, the large multiple-testing burden for genome-scale data and the necessity of a pre-screening step for genome-scale data.

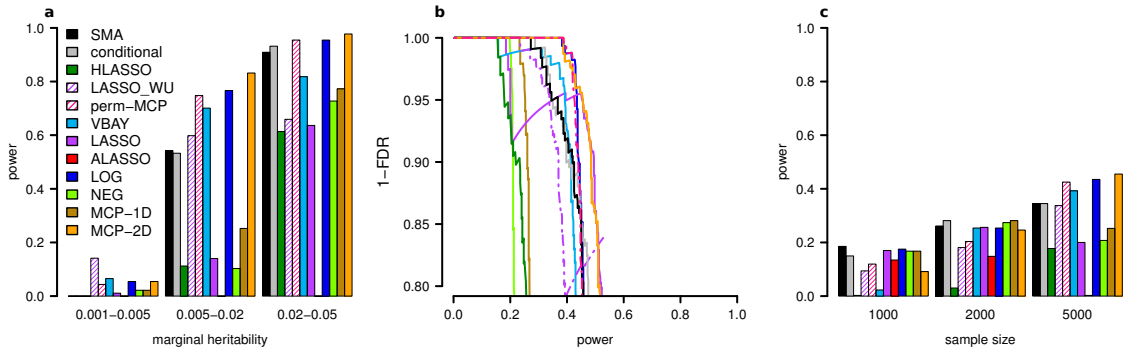
PUMA's statistical power is due to its data-adaptive properties. PUMA 1) performs a two dimensional search of the tuning parameter space 2) selects the number of nonzero coefficients based on both the fit to the data and the sample size, and 3) uses a heuristic methods to assess the significance of correlated markers. Conversely, perm-MCP fixes one of the tuning parameters, does not incorporate the sample size, and does not address the issues of testing the significance of correlated markers. Moreover, perm-MCP relies on setting the eFPR despite problem of determining an appropriate value *a priori* for high dimensional data.

2.3.5 The potential of the PMR GWAS framework as implemented in PUMA

For the 960 simulated GWAS datasets we analyzed, almost all PMR GWAS approaches implemented in PUMA except NEG and adaptive Lasso outperformed single marker analysis under simulation conditions with sufficient sample size (Figure 2.3, Figures A.3-A.10). Quite critically, the performance is far greater even when using a conservative control of FDR that is commonly employed in GWAS studies. Moreover, the improvement of the PMR methods in PUMA is most noticeable for causal variants with intermediate marginal heritability. Overall, these simulations demonstrate that the advantage of PMR methods over a single marker test increases with sample size, but decreases with the number of susceptibility loci (Figure A.3-A.4).

While the penalized methods implemented in our PUMA framework consis-

Figure 2.3: PUMA methods outperform other tests of association. Shown here are representative examples of simulation results for single marker analysis (SMA), 2-step conditional regression, a permutation based tuning of MCP (perm-MCP), our approximate Bayesian method (VBAY), and our PUMA methods (Lasso, Adaptive Lasso, LOG, NEG, 1D-MCP, 2D-MCP). Results are shown here for 20 replicate datasets from simulations with 5000 individuals, 20 causal markers affecting disease risk and a heritability of 50%. a) The power of each method to recover true associations at a fixed FDR of 5% shown as a function of the marginal heritability of each causal marker. b) Precision-Recall curve for the same simulations as in (a). Note that perm-MCP selected very few markers per simulation so the FDR did not exceed 10%. c) Power to recover true associations at an FDR of 5% for a range of sample sizes.



tently had higher power than single marker analysis as a function of FDR under most simulation conditions, none of the penalties consistently stood out as the most powerful. However, our PUMA framework, which includes a fast novel algorithm for penalized maximum likelihood estimation in generalized linear models, data-adaptive tuning of tuning parameters, heuristics for model selection and novel method of assigning p-values (see Methods) increased the power of PMR methods compared to current approaches using the same penalties [72, 209]. We note that our implementation of the NEG penalty showed a substantial increase in power over the HyperLasso program [72] and indicates that our search over tuning parameter values and heuristic approach for model

selection was successful. Moreover, our search of one or both tuning parameter values for MCP (termed 1D-MCP and 2D-MCP, respectively) showed that our approach to applying MCP (i.e. 2D-MCP) can be more powerful than that of Ayers and Cordell [9]. The fact that our implementation of Lasso had higher power than the version of Wu, et al. [209] confirms the usefulness of our data-adaptive approach for selecting penalty strength and our novel method for assigning p-values. We also note that for comparison we applied a conditional regression test and our previously published algorithm VBAY, a variational Bayes approach for fitting a mixture prior penalty [114]. We found that perm-MCP and VBAY had similar performance to our PMR methods and while the conditional test of association was sometimes more powerful than single marker analyses it was generally not as powerful as the PUMA PMR methods.

2.3.6 Summary of Wellcome Trust Case Control Consortium (WTCCC) re-analysis

In our re-analysis of type 1 diabetes, Crohn's disease and rheumatoid arthritis datasets, we applied a single-marker analysis and all PMR analysis approaches (Lasso, Adaptive Lasso, NEG, LOG, 1D-MCP, 2D-MCP, perm-MCP) using all the recommended components of our framework. We included sex and the first two principal components of the genotype matrix [150] as unpenalized covariates, applied a pre-screening cutoff of 0.01 on the p-values from the single marker test, and ran 100 reorderings for the non-convex penalties. Quantile-Quantile (QQ) plots of p-values from a standard single marker analysis indicate that the effects of any remaining population structure is minimal. Moreover,

including the subset of significantly associated markers identified by the PMR methods as covariates in a single marker analysis of remaining markers does not yield an inflation of the QQ plots and thus indicates that the PMR methods are not overfitting the data (Figure A.12). We also note that due to the complex LD around the MHC on chromosome 6, we included this region in our analysis, but we omit this region from any *post hoc* analysis and discussion.

Our single-marker re-analysis of type 1 diabetes, Crohn's disease and rheumatoid arthritis datasets reproduced the same associations as reported in the original analysis (Figure A.13). Our PMR methods recapitulated almost all of the associations identified by single marker analysis, although there were differences among the methods. The PUMA Lasso and Adaptive Lasso identified almost no additional associations compared to single marker tests, and while LOG, NEG and 1D-MCP identified more, almost all of the associations found by these five methods (Lasso, Adaptive Lasso, LOG, NEG, 1D-MCP) were identified by 2D-MCP (Figure 2.4, A.14). We note that perm-MCP identified very few associations (12 overall, across the three diseases), all but one of which was identified by a single marker test, and all were identified by 2D-MCP. We therefore discuss the associations found by 2D-MCP, where we consider three categories of interest (Table 2.3): those concordant with single marker tests, those that recapitulate associations identified in external GWAS studies but not by single marker analysis of the WTCCC, and novel associations, of which many were deemed to be biologically interpretable in terms of the current knowledge of disease etiology. In the absence of functional validation, the presence of a feasible biological interpretation lends more credibility to these novel findings.

Table 2.3: Number of associations identified in the analysis of Wellcome Trust Case Control Consortium (WTCCC) data by disease and category. Number of associations identified for Crohn's disease (CD), rheumatoid arthritis (RA) and type 1 diabetes (T1D) divided into 5 categories for the union of all associations identified by PUMA methods.

	CD	RA	T1D
Concordant with SMA	8	1	4
Replications not significant by SMA			
Independent datasets	0	0	1
Non-independent datasets	5	0	1
Etiologically relevant associations	3	1	6
Other novel associations	12	11	11

A critical point to note about the performance of our PUMA framework for PMR analysis of GWAS data is that these methods not only result in the correct identification of more loci than a single marker testing analysis (when controlling the false discovery rate at the same level), but also lead to re-orderings of the rank of markers that are considered the most significant when compared to a single marker analysis (Figure 2.5). As a consequence, we are able to identify etiologically relevant and replicated disease loci that are too weak to be detected by single marker analysis, yet show strong signals of association by PMR analysis. This means that our PMR GWAS analysis is not simply taking advantage of the lower residual variance to improve performance, but is also taking advantage of the fact that conditional correlation of a relevant marker with the phenotype is often more significant than the marginal correlation. When the coefficients for multiple markers, each tagging different susceptibility loci throughout the genome, have nonzero values in the PMR framework, their association with the phenotype becomes more significant. Our framework can therefore identify disease susceptibility loci in a GWAS with weak associations with phenotype, when they are invisible to a single marker testing approach (i.e. they have p-

values in a single marker test that would never be considered significant).

2.3.7 Associations identified by PUMA are concordant with associations from single marker tests

The associations identified by PUMA generally recapitulate associations identified by single marker analysis, and the PUMA hits have perfect concordance for strong associations. Overall 2D-MCP recapitulates the largest number of associations, while the union of the other PMR methods (considered here for illustrative purposes due to the high degree of concordance with each other, and the fact that 2D-MCP identifies almost all of the associations they find) had a lower degree of concordance with the single marker analysis (Figure 2.4, 2.6, A.14, Table A.1). Of the 6 associations identified by a single marker analysis that were missed by our methods, 5 were from type 1 diabetes and 1 was from Crohn's disease (Table A.2). One of these associations was borderline significant by 2D-MCP with a p-value of 1.42×10^{-7}.

2.3.8 PUMA methods replicate associations identified by external studies

We compared associations identified by our PUMA methods that were not detected by single marker tests in the WTCCC dataset to markers identified by independent studies in the HuGE database of published GWAS [218] in order to find associations identified in both our analysis and an independent study that

did not include WTCCC data. Such replications are considered the gold standard for validating a putative association [25]. In the ideal case the same marker would show an association in both the WTCCC dataset and those summarized in the HuGE database. However, given 1) the lack of overlap of marker-sets between genotyping platforms, 2) that the HuGE database reports only the most significant marker in an associated LD block, and 3) that PMR methods tend to select only a single marker within a LD block, we considered a marker to recapitulate a known association if the two are within 0.1 cM [10]. A representative example from Crohn's disease is shown in Figure 2.7 where only 2D-MCP is able to identify STAT3 as a susceptibility locus in the WTCCC dataset (Figure 2.7a). While this association has also been replicated in non-independent datasets [10], which included WTCCC data, the role of STAT3 in Crohn's and other autoimmune disease is well established [24, 40].

While all PUMA methods and a single marker test are able to replicate associations from independent studies, LOG, NEG and 1D-MCP, stood out in terms of identifying associations replicated by non-independent studies, but not detected in the WTCCC dataset by a single marker analysis. These counts reflect the results when the number of markers considered as 'hits' was set to be equal across methods so that they reflect the ordering of markers by PMR methods rather than the number of associations. When comparing the total number of significant hits from each method to associations identified in either independent studies or non-independent external studies that incorporated WTCCC data, 2D-MCP is the only PMR method to identify as many total replicated associations as a single marker test (Tables 2.5, A.3–A.4, Figures A.15–A.19). However, 2D-MCP is able to replicate known associations that cannot be repli-

cated by a standard single marker test in this dataset, thus demonstrating that PMR methods can extract biologically relevant information that is overlooked by standard analyses (Table A.5). These results demonstrate that PMR methods overall are able to identify replicated associations in this dataset that are invisible to a standard single marker test. Moreover, our methods provide an opportunity to replicate previously unreplicated associations by re-analyzing existing GWAS datasets.

Figure 2.4: PUMA identifies associations for Wellcome Trust Case Control Consortium (WTCCC) data that are novel and that overlap hits from previous GWAS. Genome-wide plot of associations identified by analyzing the WTCCC data for type 1 diabetes using PUMA and single marker tests. Replications from independent (not including WTCCC data) and non-independent (including WTCCC data) GWAS of the same disease are indicated with pink boxes and diamonds, respectively. For comparison, markers identified using a single marker association analysis are presented in black circles, where we note that these same hits are all identified by PUMA methods. Also for comparison, we relaxed the Bonferroni threshold for single marker analysis (open circles) until the same number of associations as found by PUMA methods are reported, where we note that many of these additional hits tend not to overlap PUMA hits or previous GWAS hits. Arrows indicate novel associations that are biologically interpretable (see Table 2.6).

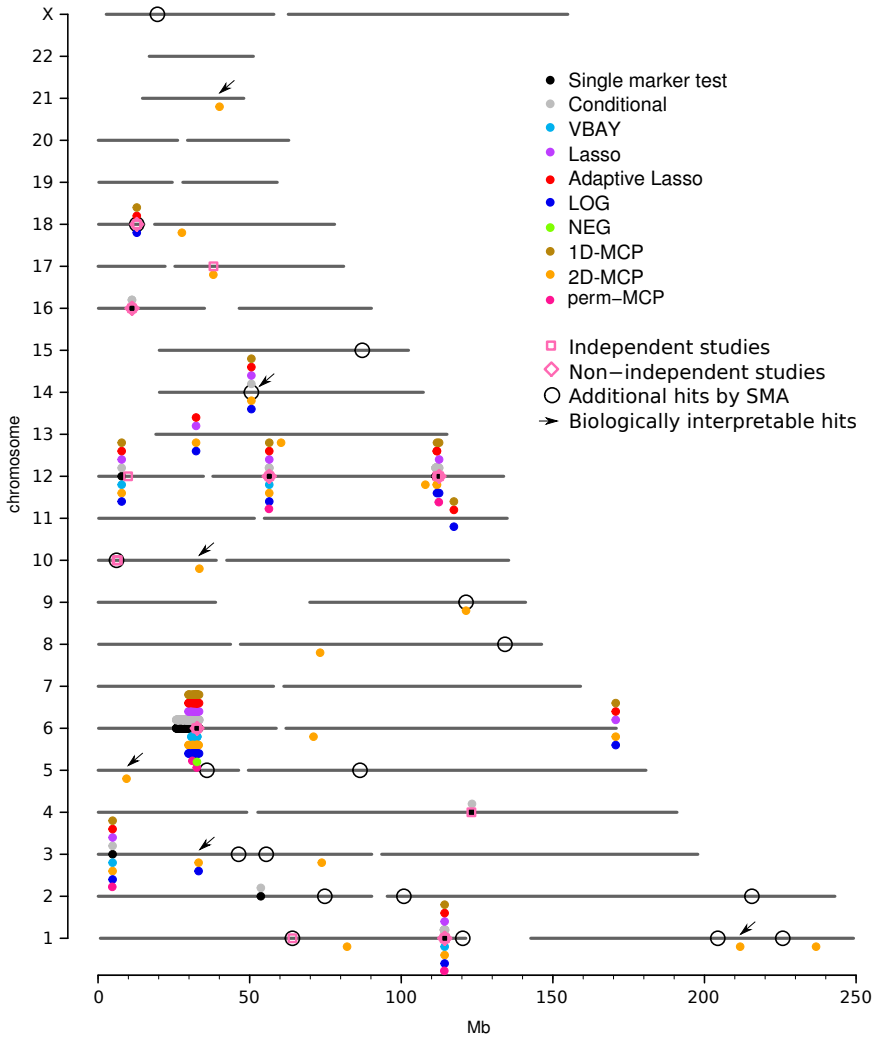


Figure 2.5: Etiologically relevant and replicated genes identified by 2D-MCP have non-significant p-values by standard single marker analysis. Quantile-quantile (QQ) plot shows results from a single marker analysis of type 1 diabetes from the WTCCC with a subset of hits identified by 2D-MCP highlighted. P-values from the single marker test are shown in black, while each orange circle indicates a region identified as significant by 2D-MCP and its location on the plot is determined by the most significant single marker analysis p-value within 0.1 cM of the significant 2D-MCP hit. Biologically relevant genes identified by 2D-MCP are shown with arrows indicating the most significant association in the region by single marker analysis. Genes shown on the left are only detectable with 2D-MCP, while genes on right are identified by both 2D-MCP and single marker analysis. P-values from the MHC region on chromosome 6 are omitted.

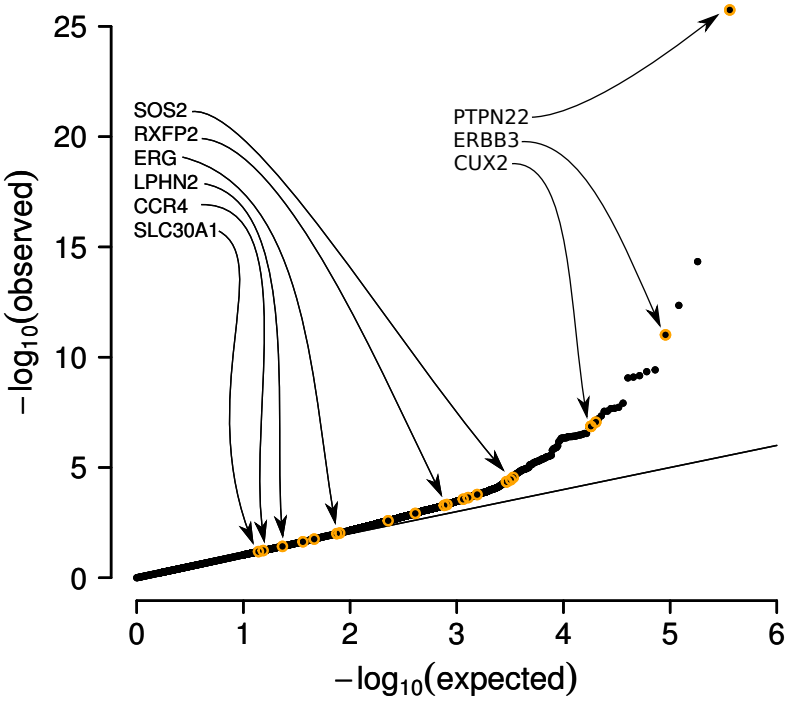


Figure 2.6: Venn diagrams showing concordance between methods.Venn diagrams show the overall concordance between regions identified by a single marker test, 2D-MCP and the union of Lasso, Adaptive Lasso, NEG, LOG, 1D-MCP and VBAY for Crohn's disease (CD), rheumatoid arthritis (RA) and type 1 diabetes (T1D) for the WTCCC analysis. Areas are approximately proportional to the counts shown and empty regions correspond to a count of zero.

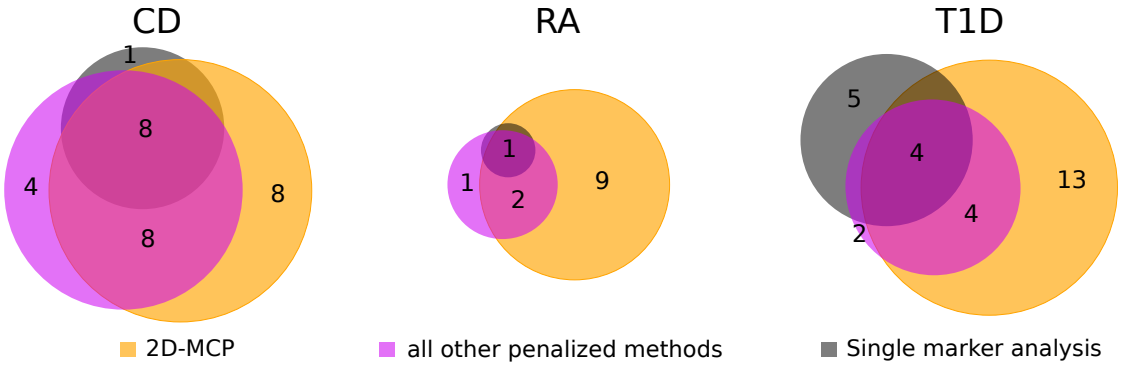


Figure 2.7: Local manhattan plots illustrating individual examples of associations identified by PUMA analysis of the Wellcome Trust Case Control Consortium (WTCCC) data. The top panel shows $-\log_{10}$ p-values (left axis, all methods except VBAY) and posterior probabilities for VBAY (right axis) for markers in the local genomic region, gene models are shown below in orange with the names of the associated gene indicated, the middle panel shows recombination rates and genetic distance from where the associated marker is indicated with an asterisk and the bottom panel shows a linkage disequilibrium plot among markers in the region using D′. a) A region identified only by 2D-MCP replicates an association from a non-independent studies (which included WTCCC data) of Crohn's disease, b) a novel association identified for type 1 diabetes only by a PUMA method (2D-MCP) that implicates the etiologically relevant SLC30A1 gene, and c) an association identified only by a PUMA method (2D-MCP) for type 1 diabetes that implicated the LPHN2, a gene previously identified but not replicated as a risk locus for type 1 diabetes. Although the associations from the independent studies do not tag the same linkage disequilibrium block as the association identified by 2D-MCP, all three likely affect LPHN2 as they are located either in or directly upstream of this gene and next closest gene is 1.8 Mb (1.7 cM) away.

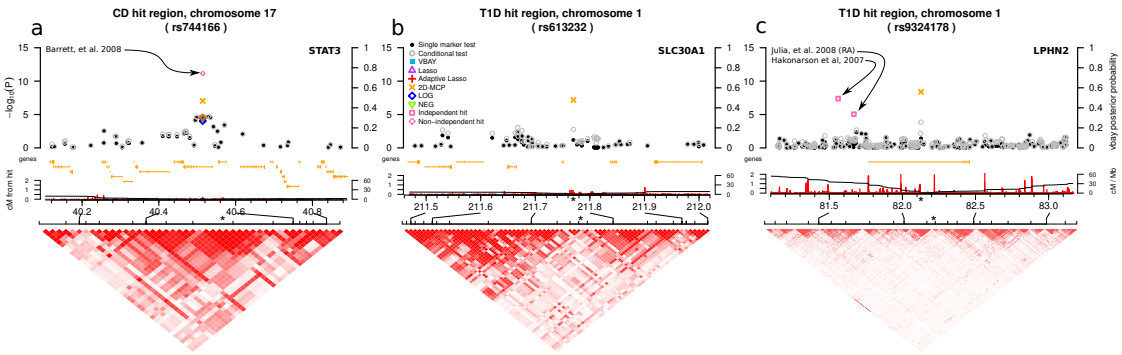


Table 2.4: Novel etiologically relevant susceptibility loci identified in Wellcome Trust Case Control Consortium (WTCCC) by PUMA methods. Genes were deemed to be etiologically relevant if they have been previously associated with etiologically related diseases or are known to function in biologically relevant pathways based on public databases and disease literature. The significance of markers with regression coefficient of exactly zero by a penalized maximum likelihood method could not be assessed and are indicated with a dash.

				Method										
disease	SNP	chromosome	position	Single marker analysis	Conditional test	VBAY	Lasso	Adaptive Lasso	LOG	NEG	1D-MCP	2D-MCP	perm-MCP	relevant genes
CD	rs903228	2p16.2	53,692,048	1.71×10^{-05}	1.71×10^{-05}	0.976	1.79×10^{-06}	1.45×10^{-06}	4.99×10^{-07}	-	7.38×10^{-06}	8.01×10^{-09}	2.58×10^{-06}	ASB3
CD	rs11627513	14q32.2	97,539,170	2.27×10^{-05}	1.25×10^{-05}	0.983	9.09×10^{-06}	1.07×10^{-05}	5.61×10^{-06}	3.02×10^{-05}	2.47×10^{-06}	1.06×10^{-09}	-	VRK1
CD	rs7497036	15q24.1	74,873,678	1.56×10^{-04}	4.3×10^{-05}	0.62	1.19×10^{-06}	1.39×10^{-06}	1.94×10^{-06}	6.38×10^{-07}	1.73×10^{-06}	6.82×10^{-07}	-	CYP11A1, SEMA7A
RA	rs12027041	1p36.32	3,591,447	7.55×10^{-06}	7.55×10^{-06}	0.0383	1.03×10^{-04}	2.83×10^{-04}	4.04×10^{-05}	-	7.4×10^{-05}	7.92×10^{-08}	-	TP73
T1D	rs613232	1q32.3	211,769,892	6.51×10^{-02}	1.71×10^{-03}	-	-	-	-	-	-	6.96×10^{-08}	-	SLC30A1
T1D	rs4074415	3p22.3	33,161,744	7.68×10^{-02}	2.17×10^{-03}	-	-	-	1.68×10^{-08}	-	-	4.45×10^{-08}	-	CCR4
T1D	rs415024	5p15.31	9,392,357	1.69×10^{-04}	1.03×10^{-04}	0.0364	2.37×10^{-05}	1.76×10^{-05}	2.1×10^{-05}	-	5.83×10^{-06}	2.59×10^{-09}	-	SEMA5A
T1D	rs9576911	13q13.1	32,329,117	1.4×10^{-03}	1.27×10^{-06}	0.103	9.75×10^{-08}	3.22×10^{-08}	6.85×10^{-07}	-	4.57×10^{-06}	9.45×10^{-08}	-	RXFP2
T1D	rs7157296	14q21.3	50,566,881	3.59×10^{-05}	1.88×10^{-07}	0.906	3.99×10^{-07}	6.79×10^{-07}	1.67×10^{-07}	-	7.72×10^{-07}	3.88×10^{-10}	-	SOS2
T1D	rs2836631	21q22.2	40,065,905	9.38×10^{-03}	4.78×10^{-04}	-	-	-	4.9×10^{-05}	-	-	1.08×10^{-08}	-	ERG

Table 2.5: Number of associations identified in re-analysis of WTCCC datasets that replicate associations from either independent or non-independent GWAS. A study is considered to be independent if it does not incorporate data from the WTCCC. An associated marker is considered to replicate a known susceptibility locus if it is within 0.1 cM a marker [10] reported as an association for the same phenotype in the HuGE database. Numbers in parentheses indicate the number of hits that are distinct from those found by the single marker test.

Independent studies									
	Methods								
disease	SMA	VBAY	SMA	Lasso	Adaptive Lasso	LOG	NEG	1D-MCP	2D-MCP
CD	6	4 (0)	6	5 (0)	4 (0)	5 (0)	5 (0)	5 (0)	5 (0)
RA	1	1 (0)	1	1 (0)	1 (0)	1 (0)	1 (0)	1 (0)	1 (0)
T1D	4	2 (0)	4	2 (0)	3 (1)	3 (1)	0 (0)	3 (1)	2 (0)
Total	11	7 (0)	11	8 (0)	8 (1)	9 (1)	6 (0)	9 (1)	8 (0)

Non-independent studies									
	Methods								
disease	SMA	VBAY	SMA	Lasso	Adaptive Lasso	LOG	NEG	1D-MCP	2D-MCP
CD	3	4 (1)	3	2 (0)	2 (0)	2 (1)	5 (3)	3 (2)	6 (3)
RA	0	0 (0)	0	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
T1D	2	0 (0)	2	0 (0)	1 (0)	1 (0)	0 (0)	1 (0)	2 (1)
Total	5	4 (1)	5	2 (0)	3 (0)	3 (1)	5 (3)	4 (2)	8 (4)
Grand total	16	11 (1)	16	10 (0)	11 (1)	12 (2)	11 (3)	13 (2)	16 (4)

2.3.9 PUMA methods identify novel associations

Re-analysis of type 1 diabetes, Crohn's disease and rheumatoid arthritis datasets from the original WTCCC [205] with our PUMA methods revealed novel associations that have not been identified in previous GWAS of these diseases (Table 2.4, Figures A.14, A.20–A.22). These methods, most notably 2D-MCP, identify novel associations in or near genes which have been previously associated with etiologically related diseases or which are known to function in biologically relevant pathways based on public databases and disease literature (Tables 2.4,2.6). In addition, PUMA also identified associations without a clear biological link to the disease phenotype (Tables A.6–A.8).

PUMA methods identified novel susceptibility loci for type 1 diabetes involved in pancreatic function, insulin pathways and immune cell function and for Crohn's disease that are involved in pro- and anti-inflammatory pathways (Table 2.6). 2D-MCP identified a gene functioning in apoptosis as a susceptibility locus for rheumatoid arthritis (Table 2.6). These genes are known to function in relevant pathways or have been previously implicated in the etiology of the disease but have not been found by previous GWAS of each disease. A representative example is shown in Figure 2.7b where only 2D-MCP identifies an association that implicates SLC30A1. This gene is a zinc transporter related to SLC30A8, which has been implicated in type 2 diabetes, and zinc transport plays a role in insulin secretion by pancreatic β-cells [86, 107].

2.4 Discussion

Each GWAS discovery that has a well supported association produces valuable information for understanding the etiology of the disease phenotype and such discoveries are regularly used as the foundation for studies that use the locus as a starting point [69, 140]. Given that GWAS involving a thousand to several thousands of individuals seldom return more than a few to a dozen well-supported associations (depending on the disease) the monetary, time, and resource investment in these studies often translates to a considerable expenditure per discovery. This is true even when considering additional discoveries that may occur as individual GWAS are combined together into large meta-analysis studies [3, 10, 38, 46, 49, 95, 176]. We have demonstrated that our PUMA framework has the potential to produce added investment return for

Table 2.6: Novel susceptibility loci identified by PUMA methods and their biological link to the disease. Genes were deemed to be etiologically relevant if they have been previously associated with etiologically related diseases or are known to function in biologically relevant pathways based on public databases and disease literature.

disease	gene	description
CD	ASB3	functions in ubiquitination and degradation of TNF-R2, which mediates TNF-α pro-inflammatory response [29, 212]
CD	VRK1	phosphorylates c-Jun and p53, which both function in inflammation [171]
CD	CYP11A1	cytochrome P450 enzyme that synthesizes anti-inflammatory corticosterone in the intestine, and the enzyme is underexpressed in inflamed colon biopsies of patients with Crohn's disease [31, 137]
CD	SEMA7A	immune semaphorin whose expression on activated T-cells induces macrophage production of pro-inflammatory cytokines [181]
RA	TP73	p53-like transcription factors that functions in apoptosis, a process implicated in the etiology of rheumatoid arthritis [110, 213]
T1D	SLC30A1	zinc transporter related to SLC30A8, which has been implicated in type 2 diabetes, and zinc transport plays a role in insulin secretion by pancreatic β-cells [86, 107]
T1D	CCR4	chemokine receptor and CCR4-bearing T-cells function in the autoimmune inflammation of the pancreas in mice [89]. A nearby marker shows a strong association with celiac disease [36]
T1D	SEMA5A	member of the semaphorin protein family whose members play a role in cell-cell interactions in immune processes, but the function of this gene is not well characterized [181]
T1D	RXFP2	receptor for relaxin, a member of the insulin protein family [61]
T1D	SOS2	Ras-guanine nucleotide exchange factor which is upstream of a number of relevant signalling pathways [132]
T1D	ERG	ETS-family transcription factor that functions in pancreatic development [90]

GWAS studies by discovering additional well-supported disease loci associations that are invisible to the standard single marker analysis methods responsi-

ble for almost all reported GWAS [68, 218]. For example, our re-analysis of type 1 diabetes, Crohn's disease and rheumatoid arthritis from the original Wellcome Trust Case Control Consortium (WTCCC)[205] demonstrates that PUMA methods can identify associations that are not detectable by single marker analysis approaches but which replicate associations known from independent studies, which did not include WTCCC data, as well as novel loci with strong links to known disease etiology. These included 10 novel associations identifying genes that are linked to primary pathways of these autoimmune diseases, specifically 6 genes involved in pancreatic function, insulin pathways and immune-cell function in type 1 diabetes; 4 genes (in 3 association regions) functioning in pro- and anti-inflammatory pathways in Crohn's disease; and 1 gene involved in apoptosis pathways in rheumatoid arthritis. Applying our PUMA framework therefore has the potential to add a significant number of discoveries for a given GWAS.

A critical property of our PUMA framework is it does not return the same ordering of significant markers produced by a standard single marker analysis. By simultaneously accounting for the associations of multiple loci and better reflecting the underlying polygenic architecture of complex phenotypes, PUMA can find strong statistical support for associations deemed non-significant by a single marker analysis and places these among the top list of associations. A prime example is marker rs613232 which had a p-value of 6.51×10^{-2} by a single marker analysis in the type 1 diabetes dataset so it would not be considered for a follow-up study. However, by taking into account the polygenic architecture of the trait, 2D-MCP assigned it a p-value of 6.96×10^{-8} (Figure 2.7b, Table 2.4). This marker tags the zinc transporter SLC30A1 and zinc transport has an

established role in type 1 diabetes, yet this gene was only identified as a susceptibility locus by 2D-MCP. This example illustrates the power of PUMA methods to reorder the p-values of markers so that a marker that is not in the top 20,000 by a single marker test can be in the top 30 by 2D-MCP. Another example is that of LPHN2, a gene identified by an independent GWAS of type 1 diabetes, yet the association was not replicated in an independent dataset in the same study [59] or, to our knowledge, any subsequent study. In our re-analysis, 2D-MCP identified a strong signal in a nearby marker and assigned it a p-value of 3.99×10^{-9}, while the p-value by a single marker test was 3.78×10^{-2} (Figure 2.7c, Table 2.4). The very weak single marker p-value found in this dataset makes the previous inability to replicate this association unsurprising. The gene encodes the G-coupled protein receptor latrophilin 2 and has a weak association with rheumatoid arthritis [81] but its relation to disease etiology is unclear. These examples illustrate that our PUMA framework returns additional and complimentary information to the results of a single marker analysis of a GWAS. In general, it seems clear that applying a spectrum of appropriate GWAS analysis methods to the same data is likely to maximize discovery.

The PUMA framework and software that we present here is immediately applicable to a large number of existing GWAS and we are currently exploring extensions of the framework to address additional challenges in GWAS experimental designs and GWAS analysis. For example, GWAS discoveries are regularly being produced by consortia that combine several independently executed GWAS experiments. Such combined data introduce a number of complexities including complex batch effects, population structure, relatedness, and latent environmental variables. While meta-analysis techniques for combining

p-values across studies are a good approach to normalizing for many of these issues [3, 10, 38, 46, 49, 95, 176], a PMR analysis directly on the genotypes can include correction for study heterogeneity, population structure and cryptic relatedness using a linear mixed model [108, 169, 170], and we are currently working on such extensions. Given that the increase in performance for our PMR methods compared to single marker analysis increases with increasing sample sizes, solving these problems has great potential to detect additional weak associations. There is also going to be a near-term shift towards GWAS that add millions of additional genetic markers genotyped by next-generation sequencing, which can add increased density of markers and different allele types. Our approach can already handle these large number of markers directly to take advantage of the better tagging, and in some cases genotyping, of causal disease polymorphisms. The trend of increased sample sizes and genome marker coverage in GWAS also opens the opportunity to identify genetic interactions that are currently difficult to detect, including epistasis and gene$\times$environment interactions, which could be identified by incorporating group penalty approaches [129, 219, 227] within our framework. Overall, our framework represents a platform for integrating richer statistical models and techniques for addressing the future needs of GWAS.

2.5 Methods

2.5.1 The PUMA framework

Our framework is a combination of algorithms and heuristic approaches designed for robust and efficient analysis of GWAS datasets when the desired output is a ranked list of genetic markers that individually tag disease loci. The value of the framework is that tag genetic markers, which are too weak to be reliably identified by a single marker analysis, can be identified while preserving a conservative FDR genome-wide. To solve issues that have limited the value of existing PMR GWAS software for this purpose, we designed our framework to have the following properties: 1) the versatility to handle a diversity of penalties for simultaneous analysis of thousands to millions of genetic markers while incorporating unpenalized covariates, 2) the efficiency to analyze up to millions of markers after pre-screening on a standard desktop, 3) the sensitivity to tune the strength of penalties and to perform model selection when the fraction of variation accounted for by disease loci identifiable with tag markers is small (as is typical for GWAS), and 4) the capability to return a ranked list of p-values where each of the top markers identifies an independent disease association. We outline the components of our framework responsible for each of these properties in the next four sections, followed by a description of our software PUMA that implements our recommended practices and options for implementation. We note that in its entirety, this framework is a new GWAS analysis approach that incorporates novel components including (but not limited to): application of penalties not previously applied to GWAS, a new MM algorithm for GLMs, heuristics for penalty strength and model selection, and *post hoc* model fitting

approaches for ranking associated markers.

2.5.2 Objective functions and penalties

Our framework makes use of a generalized linear modeling (GLM) framework to construct the likelihood objective function. We can therefore model phenotypes measured on a large diversity of scales by implementing an appropriate link function [125], although here, we limit our implementations to an identity and logistic link function to model continuous phenotypes with normal error and case-control phenotypes, respectively. We also note that incorporating unpenalized covariates is straightforward, where these are modeled with regression coefficients with no penalty. While our current implementation is restricted to penalties that select a small number of well supported markers (i.e. feature selection penalties [42, 56, 124, 187, 221, 232]), the framework is versatile enough to implement a wide diversity of penalties approaches when making use of the algorithm described in the next section.

For marker selection, we use the penalized maximum likelihood estimate (pMLE) of the regression coefficients:

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmax}} \ \ell(\mathbf{y}|\boldsymbol{\beta}) - \sum_j p_\theta(\beta_j) \tag{2.1}$$

where $\mathbf{y}$ is the vector of disease phenotype values, and $\boldsymbol{\beta}$ is the vector of regression coefficients $\ell(.)$ is the log-likelihood of a linear or logistic regression and $p_\theta(.)$ is the penalty function on the magnitude of β_j indexed by a vector of tuning parameters, θ. Since we are interested in identifying a small set of variables associated with the phenotype, the penalty function must have the

sparsity property whereby most of the regression coefficients are set to exactly zero. Multiple penalties satisfy this condition while balancing computational tractability with desirable theoretical properties. We implement the penalties that have been applied for PMR GWAS analysis (i.e. Lasso, Adaptive Lasso, NEG, MCP) as well as a penalty that has not be previously applied to GWAS (i.e. LOG). We describe the properties of each penalty in the following paragraphs and the functional form of each penalty is given in Appendix A.1.

The Lasso penalty [187] (previously implemented for PMR GWAS in the software Mendel [209]) is a linear function of the magnitude of the regression coefficients and is the most widely used since it has a single tuning parameter. Moreover, the penalty is convex so that it yields a unique pMLE. Yet it is widely known to select too many variables with non-zero coefficients in high dimensional datasets [42] and does not satisfy the "oracle property" whereby parameter estimates are asymptotically equivalent to unpenalized estimates when the identity of the non-zero coefficients are known in advance [42].

The Adaptive Lasso penalty [232], (previously implemented for PMR GWAS by Yang, et al. [214]) unlike the Lasso, satisfies the oracle property. This two-step Lasso regression procedure is also convex (yields a unique pMLE) although it requires an initial estimate of the regression coefficients, which are then used to weight the strength of a Lasso penalty in the second step. There is no criterion for determining optimal weights, so in practice the Lasso penalty for each coefficient is weighted by the square root of the initial coefficient estimate.

The NEG penalty [56] (previously implemented for PMR GWAS in the soft-

ware HyperLasso [72]) has two tuning parameters and is non-convex such that it produces a multimodal likelihood surface where pMLE's are not unique. The penalty satisfies the oracle property, since the derivative of the penalties approach zero in the limit [42], although it has other less desirable properties since its derivative approaches zero much more slowly than the other penalties and its very complex functional form makes it numerically unstable for large coefficient values. In our framework, we re-implement NEG using a faster algorithm than Hoggart et al. [72] and includes a two dimensional search of the tuning parameter space, while [72] use asymptotic theory to set the tuning parameters.

The MCP penalty [221] (previously implemented in the R package *grpreg* [15]) has two tuning parameters. Like NEG, this penalty is non-convex and satisfies the oracle property. However, the derivative of MCP reaches zero for finite coefficient values so that it avoids over penalizing large coefficient values. Moreover, MCP is designed to reduce the multimodality of the objective function [221]. The tuning parameters determine the slope of the penalty near the origin (i.e. λ) and coefficient value at which the derivative of the penalty is set to zero (i.e. a or γ, depending on notation). When applying this method to GWAS data, Ayers and Cordell [9] fixed the value of a at 30, and identify the value of λ using a permutation approach. We term this approach perm-MCP. In addition, we consider a one dimensional search over the value of λ and a two dimensional search over both parameters, termed 1D-MCP and 2D-MCP, respectively. The latter has the most potential since it explores the value of the tuning parameter, a, that determines the coefficient value at which the derivative of the penalty is set to zero, and learns the value of the parameter based on the data.

We also implement the LOG penalty and apply it to GWAS for the first time. The LOG penalty [124] has two tuning parameters and is non-convex, such that it produces a multimodal likelihood surface where pMLE's are not unique, but is satisfies the oracle property, since the derivatives of the penalties approach zero in the limit [42]. This penalty is also designed to identify fewer non-zero regression coefficients.

2.5.3 Minorize-Maximization (MM) algorithm and scalable implementation

Our framework implements a highly efficient algorithm and optimized coding practices to allow fast simultaneous analysis of genetic markers in the range of hundreds of thousands to millions. We implement a new minorize-maximization (MM) algorithm for finding pMLE by using a coordinate-wise ascent approach with an upper bound on the second derivative of the likelihood function [75]. By using bounded univariate updates, the algorithm is extremely fast and is guaranteed to converge to a mode of the likelihood surface. While Newton-Raphson algorithms must evaluate the log-likelihood after each update to check if it has decreased [14], the use of an MM algorithm for logistic regression guarantees a monotonic increase and eliminates the expensive function evaluation. Derivations are given in Appendix A.1 . In addition to the algorithm, we also implement a number of optimized coding practices to accelerates these PMR methods. These include data structures to minimize access time to each marker, use of optimized linear algebra libraries and searching multiple modes of the likelihood surface for non-convex penalties in parallel.

2.5.4 Adaptive tuning of penalties and model selection

Critical to the performance of our framework is preserving a conservative control of the FDR for identified markers. To accomplish this, we employ a strategy that allows our PMR methods to automatically adapt, not only to the sample size of the dataset, but also to the number and magnitude of the non-zero regression coefficients for relevant markers associated with the phenotype. Our approach includes an adaptive tuning of penalty strength in combination with model selection and assessment of model fit.

Statistical theory considering linear regression has shown that for a sample size of n, the number of variables detectable as having nonzero coefficients is on the order of $\sqrt{n}$ [222]. This is consistent with other theoretical work [74, 148] and satisfies our intuition that the number of detectable associations is directly related to the sample size of the dataset. For adaptive tuning of the Lasso and Adaptive Lasso, where the likelihood is convex and there is a single tuning parameter, the search is simple and we start with a severe penalty, which is gradually decreased to select one additional non-zero coefficient at a time, until $1.5\sqrt{n}$ genetic markers are selected. For the non-convex penalties [9, 72, 124], a grid search over a two-dimensional space of tuning parameters is used starting from equally spaced Lasso models (a special case of all non-convex penalties) where the non-convexity of the penalty is gradually increased until $1.5\sqrt{n}$ markers are selected. This approach for searching the space has been shown to avoid some suboptimal modes of the likelihood surface [124]. We note that we have previously published the approximate Bayesian methods, VBAY and VBAYNET, which incorporate a probabilistic bound, where we applied the same $1.5\sqrt{n}$ bound [114, 115]. In order to mitigate the problem of suboptimal modes

at least to some degree, we explore multiple modes of the likelihood surface for non-convex penalties by permuting the order in which the regression coefficients are updated. For both the simulations and WTCCC analyses of this study, we found 100 reorderings was sufficient to obtain robust results.

Once sets of markers with nonzero coefficients are identified for each value of the tuning parameters for a given penalty, the optimal set is determined. We assessed the overall appropriateness of the fit of a selected model based on a QQ plot by fitting an unpenalized model with selected markers, and calculating p-values for each marker in the dataset by regressing it against the residuals from the first step (Figure A.12) [113].

2.5.5 *Post hoc* assessment of p-value ranks

The most valuable final output of a PMR GWAS analysis is a ranked list of markers in decreasing order of highest confidence. Standard methods for producing such ranked lists in a PMR framework assess significance by conducting variable selection on a subset of the data and assessing significance on another subset [131], or subsetting the data many times and identifying variables selected in many of the subsets [130]. Such methods can be very computationally demanding for large GWAS datasets and have been shown to underperform a standard single marker test of association [2]. Moreover, these methods do not address the challenging problem of assessing the significance of a marker in the presence of correlated markers within the same linkage disequilibrium (LD) block. While PMR methods tend to select a single non-zero regression coefficient for an associated LD block, cases often arise where multiple markers in a LD block

have non-zero coefficients. In such cases, the correlation between the markers in the block dramatically increases the sample variance of the coefficients so that the markers are not assigned significant p-values even if each would be significant if the other were dropped from the model.

We deal with the problem of producing an informative ranked list of selected markers in our framework by initially fitting an unpenalized regression model with all markers selected by a given PMR and including relevant covariates, and calculating p-values for each marker using a standard likelihood ratio test by comparing the full model to null models where each marker is omitted in turn. The correlation between all pairs of selected markers is then evaluated, the pair of markers with the largest correlation is identified, the marker with the smallest absolute regression coefficient of the two is dropped, and p-values are then recalculated for the remaining markers. This process is repeated until no pairwise correlation between remaining markers exceeds 0.1 and each marker is finally assigned the smallest p-value produced for it during this processes. This heuristic procedure means that the values cannot be interpreted strictly as asymptotic p-values [209], but they can be considered as scores indicative of the significance of the association that ranks the values in terms of confidence while ensuring at least one marker in the LD block is assigned a p-value rank that can appropriately reflect a true association.

2.5.6 PUMA software

PUMA implements both linear and logistic models for PMR methods, as well as single marker analysis, a conditional regression analysis, and the variational

Bayes multiple regression method VBAY [114]. The software reads genotype files in TPED format and phenotypes in TFAM as used by Plink [154]. For all multiple marker methods, we employ a heuristic to remove markers with low marginal correlations with the phenotype as they are extremely unlikely to be selected to have nonzero regression coefficients [43, 209]. This is approach is not novel [43, 209], but accelerates computation and allows flexibility when analyzing extremely large GWAS datasets. The set of markers identified at each mode of the likelihood surface is stored and is saved in a text file readable by R. The software is available at `http://mezeylab.cb.bscb.cornell.edu/Software.aspx`.

For the single marker analysis, p-values are calculated using a standard F-test or likelihood ratio test [94] for linear and logistic models, respectively. The conditional regression performs the standard single marker test of association and includes the significant markers as covariates in a second set of single marker tests. The minimum of the two p-values from the first and second stage analysis is then reported for each marker. In this analysis we used a first stage p-value cutoff of 1×10^{-6} and selected the single most strongly associated marker within 100 Kb to include as covariates in the second stage. We also re-implemented the approximate Bayesian method VBAY, previously developed by our group [114]. While Bayesian regularized regression methods using a number of prior distributions as been applied to association mapping using Markov chain Monte Carlo (MCMC) methods, these cannot simultaneously analyze more than a few hundred to a few thousand genetic markers at a time [57, 103]. VBAY [114] uses a variational Bayes approximation to the posterior surface [197] and applies a hierarchical mixture prior on the regression coefficients so that the large

majority of coefficients have a high posterior probability of being exactly zero (see Logsdon et al. [114] for a more detailed discussion of this method). Within PUMA, we re-implemented VBAY and added the capability to analyze case-control phenotypes, where we increased the power to detect weak associations for case-control phenotypes by approximating a logistic regression by modeling the error distribution with a Student t-distribution with 7.3 degrees of freedom. This parameterization has the smallest squared error loss of any t-distribution with respect to the logistic error function [1, 138]. Moreover, we address the multimodality of the posterior surface by exploring many posterior modes and applying Bayesian model averaging [71] in order to weight the contribution of each mode to the posterior probability of association for each marker. The VBAY algorithm was run with 1000 restarts to explore the non-convex posterior surface. VBAY reports the posterior probability, between 0 and 1, that each marker is associated with the phenotype.

2.5.7 PUMA software recommended usage

The default settings of PUMA are our recommended settings for a GWAS analysis, which are the same procedures we followed for our analysis of the WTCCC data here:

1. Marker imputation: Beagle [18] was used to impute missing genotypes, but other methods can be used. Alternatively, PUMA fills in missing data with the mean of each marker.
2. Inclusion of fixed covariates: Identify relevant covariates and principal components and perform single marker analysis so that the correspond-

ing QQ plot and λ_{GC} [34] are acceptable. Results from the single marker analysis must be acceptable before applying PMR methods.

3. Marker filtering: We applied a pre-screening filter based on p-values from single marker analysis using a cutoff of 0.01.

4. Number of restarts: The penalized likelihood for the 1D-MCP, 2D-MCP, NEG and LOG penalties is non-convex so 100 reorderings were used to explore the multimodal surface for each setting of the tuning parameters.

5. Performance assessment: We recommend assessing the fit of a PMR model by including the selected markers as covariates in a subsequent single marker analysis. Too much inflation or deflation of the p-values indicates that the PMR method may be overfitting the data.

6. Threshold determination: We have demonstrated that the reported p-value score statistics are valuable at prioritizing the top hits as well as novel weak associations, so assessing the list in rank order is the suggested strategy for minimizing false positives. For example, in our current analyses we examined at most the top 30 hits for each method combined across the three diseases, which limited our focus to markers with a p-value score of $< 1 \times 10^{-7}$ for 2D-MCP, $< 1 \times 10^{-6}$ for all other PMR methods, and for comparison, a posterior probability of > 0.97 for VBAY.

2.5.8 GWAS simulation study

Our approach was to simulate different sized GWAS experiments where we used the real genetic markers for unrelated European individuals from the Multi-Ethnic Study of Atherosclerosis (MESA) [13] genotyped on the

Affymetrix 6 platform. Larger sample sizes were generated by drawing haplotypes from existing individuals in order to avoid the confounding effect of population structure [150]. For each simulated GWAS dataset, we considered different sample sizes ($n = 1000, 2000, 5000$) with equal numbers of case and control phenotypes simulated under an additive threshold model with a disease prevalence of 50%, using the GCTA program [216], and that different numbers of susceptibility loci ($q = 10, 20, 30, 50$) contributed to phenotype heritability, where the total contribution of these loci to heritability was varied ($h^2 = 30, 40, 50, 60\%$). Coefficients were drawn from a $\Gamma(1, 1)$, independent of allele frequency, so that most effect-sizes were very small as determined by the marginal heritability calculated by GCTA [216]. We considered 20 replicates per simulation condition to give 960 simulated GWAS datasets. Causal variants affecting the phenotype were selected uniformly from the set of genetic markers with minor allele frequency (MAF) $> 5\%$. We followed typical array-based GWAS by omitting the causal variants from the analysis so a susceptibility locus must be identified by markers in linkage disequilibrium with the causal variant.

Following the performance evaluation of previous studies [9, 72], a marker was considered a true positive hit if it had $r^2 \geq 0.05$ with a causal marker, otherwise it was considered a false positive hit. Since a causal variant may be tagged by multiple true positive markers, the true positive count is defined as the total number of causal variants tagged when all true positive hits are considered together. Alternatively, since false positive hits will often fall in clusters in the same linkage disequilibrium block, we assign each to a 100 kb cluster centered at the most significant hit in that cluster. The false positive count is then defined as the number of such false positive clusters. We note this is a strategy for

assessing the performance properties that will be of greatest interest to GWAS practitioners since it focuses on correct identification of tag markers that are in high linkage disequilibrium (LD) and in close physical proximity to the location of the true causal alleles, while considering a strict control of the FDR.

2.5.9 Analysis of WTCCC data

To run the data analysis we used the same quality control filters as in the Wellcome Trust Case Control Consortium, first by excluding 809 individuals because of poor sample quality, non-Caucasian ancestry, or a high degree of relatedness [205]. An additional individual was removed for being an outlier by principal components analysis [150]. Marker locations and genetic map are based on reference assembly GRCh37/hg19 and dbSNP v131. Next, the same study-wide missing data rate and deviation from Hardy-Weinberg equilibrium cut-offs were used for each set of cases as in the original study [205], with an additional filter to only include markers in the analysis with a minor allele frequency greater than 0.05 in each combined case-control population, leaving approximately 360,000 markers for each combined case-control data set. We used the CHIAMO calling scores to set data to missing, where any call with a score of less than 0.90 was set to missing [205]. To impute this sporadic missing data we used Beagle [18], with the default settings and allocating a maximum of 3000 MB of memory, where the sporadic missing data for each cohort was imputed separately. The same set of controls (1958 Birth Cohort (58C) and UK Blood Service sample (NBS)) were used for each set of cases as in the original study [205]. Finally, the PMR and single marker analyses included sex as a covariate along

with the first two principal components of the genotype matrix.

In order to explore the biological function, relevant pathways and possible disease implications of each gene near a significantly associated marker, we mined public databases including GenBank [12], Pfam [153], KEGG [82], OMIM [62], GeneCards [177] as well as the HuGE database [218] of known GWAS hits and known gene-phenotype links. We also conducted an extensive literature search with each gene name and relevant phenotypes.

CHAPTER 3

CORRECTING FOR POPULATION STRUCTURE AND KINSHIP USING THE LINEAR MIXED MODEL: THEORY AND EXTENSIONS

3.1 Abstract

Population structure and kinship are widespread confounding factors in genome-wide association studies (GWAS). It has been standard practice to include principal components of the genotypes in a regression model in order to account for population structure. More recently, the linear mixed model (LMM) has emerged as a powerful method for simultaneously accounting for population structure and kinship. The statistical theory underlying the differences in empirical performance between modeling principal components as fixed versus random effects has not been thoroughly examined. We undertake an analysis to formalize the relationship between these widely used methods and elucidate the statistical properties of each. Moreover, we introduce a new statistic, effective degrees of freedom, that serves as a metric of model complexity and a novel low rank linear mixed model (LRLMM) to learn the dimensionality of the correction for population structure and kinship, and we assess its performance through simulations. A comparison of the results of LRLMM and a standard LMM analysis applied to GWAS data from the Multi-Ethnic Study of Atherosclerosis (MESA) illustrates how our theoretical results translate into empirical properties of the mixed model. Finally, the analysis demonstrates the ability of the LRLMM to substantially boost the strength of an association for HDL cholesterol in Europeans.

3.2 Introduction

Population structure and kinship are widespread confounding factors in genome-wide association studies (GWAS) that can decrease power and increase the false positive rate of tests of association [152]. As a result, it is common practice to infer population structure and kinship based on genome-wide SNP data and to exclude problematic individuals or account for these effects in the test of association [152]. Principal components analysis (PCA) is widely used to detect population structure [144]. The inferred principal components capturing the genetic ancestry of each individual are often included as fixed effects in a regression-based test of association in order to account for population structure [150, 152]. More recently, a linear mixed model (LMM) that considers the genome-wide similarity between all pairs of individuals was proposed to account for population structure, known kinship as well as cryptic relatedness [83, 84], and recent technical advances have made such models tractable for very large datasets [83, 108, 146, 182, 230].

While simple tests of association assume statistical independence between individuals, population structure and kinship indicate covariance between individuals based on the genetic similarity between individuals and the heritability of the phenotype [83]. Since it is well established that ignoring this covariance in a test of association produces deflated p-values that do not follow a uniform distribution under the null [34], it is common to apply a LMM or include principal components as fixed effects in order to model the dependence structure [152]. Both approaches model this covariance between individuals, and both can be stated as regressing the phenotype on principal components of the genotype

matrix [8, 79, 216] so that the LMM essentially includes principal components as a random effect rather than a fixed effect. While the top principal components capture population structure, explicitly modeling the pairwise relatedness between all individuals captures both population structure and kinship [83, 87, 152, 207]. Thus much recent attention has focused on the LMM since it shows better empirical performance in modeling the dependence structure of GWAS datasets [83, 87, 152, 207].

Motivated by the empirical differences between the LMM and including principal components as fixed effects, we describe a unified framework that connects these models. This framework facilitates a statistical examination of the methods' differing frequentist vs. Bayesian interpretations, their differing approaches to inference and how these differences drive their empirical properties. We next introduce a summary statistic, the effective degrees of freedom, that measures overall model complexity and the influence of each principal component on the fit of the LMM. Leveraging the unified framework and the effective degrees of freedom, we propose a novel method, the low rank linear mixed model (LRLMM) using the algorithm of Lippert et al. [108], that learns the dimensionality of the correction for population structure and kinship.

3.3 Methods

3.3.1 Modeling principal components as fixed versus random effects

Considering the matrix of genotype data $\mathbf{X}$ $(n \times p)$ for n individuals and p genetic markers, where entry $\mathbf{X}_{k,j} \in \{0, 1, 2\}$ represents the number of copies of the minor allele that individual k has of marker j, the singular value decomposition underlying principal components analysis (PCA) has the form

$$\mathbf{X} = \mathbf{USV}^T \tag{3.1}$$

so that the first i principal components are the first i columns of $\mathbf{U}$ $(n \times n)$, $\mathbf{S}$ $(n \times n)$ is diagonal so that $\mathbf{S} = \text{diag}(\mathbf{s})$ where $\mathbf{s}$ contains singular values corresponding to each principal component, $\mathbf{V}$ $(p \times n)$ stores the loadings on each marker, and each marker in $\mathbf{X}$ has been mean centered and scaled [144]. Including the first i principal components as fixed effects in a linear model takes the form

$$\begin{aligned} \mathbf{y} &= \mu + \mathbf{x}_j\beta + \mathbf{U}_{1:i}\boldsymbol{\omega} + \boldsymbol{\epsilon} \\ \boldsymbol{\epsilon} &\sim \mathcal{N}(0, \sigma_e^2) \end{aligned} \tag{3.2}$$

where $\mathbf{y}$ $(n \times 1)$ is a vector of phenotype values, μ is the scalar mean term, $\mathbf{x}_j$ $(n \times 1)$ is the j^{th} marker with scalar regression coefficient β, $\mathbf{U}_{1:i}$ are the first i principal components with coefficient vector $\boldsymbol{\omega}$ $(i \times 1)$, and $\boldsymbol{\epsilon}$ $(n \times 1)$ is the normally distributed residual error term with variance σ_e^2. Principal components are treated as fixed effects, such that maximizing the likelihood involves directly estimating all parameters. From a Bayesian perspective, the model does

not have an explicit prior on regression coefficients, $\boldsymbol{\omega}$, and thus implies a uniform prior. Furthermore, scaling each principal component by any value yields a statistically equivalent model with respect to the genetic term, $\mathbf{x}_j\beta$, since the prior on the coefficients, $\boldsymbol{\omega}$, is implicitly uniform.

Now consider the linear mixed model (LMM)

$$\begin{aligned} \mathbf{y} &= \mu + \mathbf{x}_j\beta + \boldsymbol{\alpha} + \boldsymbol{\epsilon} \\ \boldsymbol{\alpha} &\sim \mathcal{N}(0, \mathbf{K}\sigma_a^2) \\ \boldsymbol{\epsilon} &\sim \mathcal{N}(0, \sigma_e^2) \end{aligned} \tag{3.3}$$

where $\boldsymbol{\alpha}$ $(n \times 1)$ is a random effect vector with a multivariate Gaussian prior, $\mathbf{K}$ $(n \times n)$ is the genetic similarity matrix between all pairs of individuals so that $\mathbf{K}_{k,l}$ represents the similarity between individuals k and l, and σ_a^2 is the additive genetic variance. Here population structure is treated as a random effect and fitting the model involves integrating over the vector $\boldsymbol{\alpha}$ with respect to the Gaussian prior so that the likelihood is maximized with respect to σ_a^2, σ_e^2, μ, and β [84, 172].

For simplicity, let the genetic similarity matrix $\mathbf{K}$ be a simple function of observed genotypes as in Patterson et al. [144], and consider the singular value

decomposition from equation (3.1) and the factorization of $\mathbf{K}$

$$\begin{aligned}
\mathbf{K} &= \mathbf{X}\mathbf{X}^T \\
&= \mathbf{U}\mathbf{S}\mathbf{V}^T(\mathbf{U}\mathbf{S}\mathbf{V}^T)^T \\
&= \mathbf{U}\mathbf{S}\mathbf{V}^T\mathbf{V}\mathbf{S}\mathbf{U}^T \\
&= \mathbf{U}\mathbf{S}\mathbf{V}^{-1}\mathbf{V}\mathbf{S}\mathbf{U}^T \\
&= \mathbf{U}\mathbf{S}^2\mathbf{U}^T \\
&= \mathbf{U}\mathbf{S}(\mathbf{U}\mathbf{S})^T \\
&= \mathbf{R}\mathbf{R}^T
\end{aligned} \tag{3.4}$$

so that the columns of $\mathbf{U}$ are the principal components of the genotype matrix, $\mathbf{X}$, exactly as in equation (3.1), and, by construction, the columns of $\mathbf{R}$ $(n \times n)$ are the principal components weighted by their respective singular values. We note that each principal component $\mathbf{U}_t$ has a singular value s_t and eigenvalue s_t^2. Using the property of a multivariate Gaussian that $\boldsymbol{\phi} \sim \mathcal{N}(\mathbf{m}, \boldsymbol{\Sigma}) \implies \mathbf{B}\boldsymbol{\phi} \sim \mathcal{N}(\mathbf{B}\mathbf{m}, \mathbf{B}\boldsymbol{\Sigma}\mathbf{B}^T)$, and the decompositions in (3.4), it is apparent that $\boldsymbol{\gamma} \sim \mathcal{N}(0, \sigma_a^2) \implies \mathbf{R}\boldsymbol{\gamma} \sim \mathcal{N}(0, \mathbf{K}\sigma_a^2)$, so the LMM (3.3) can be rewritten equivalently as

$$\begin{aligned}
\mathbf{y} &= \mu + \mathbf{x}_j\beta + \mathbf{R}\boldsymbol{\gamma} + \boldsymbol{\epsilon} \\
\boldsymbol{\gamma} &\sim \mathcal{N}(0, \sigma_a^2) \\
\boldsymbol{\epsilon} &\sim \mathcal{N}(0, \sigma_e^2)
\end{aligned} \tag{3.5}$$

Based on the relationship between equations (3.2) and (3.5), it is apparent that modeling principal components as fixed or random effects share the same underlying regression model. This transformation explicitly formalizes the previously described relationship between modeling principal components as fixed versus random effects [8, 79, 216]. While the LMM includes all principal components, only $i \ll n$ principal components are included in the fixed effects model

since the number of covariates cannot be on the same order as the sample size while still maintaining reasonable statistical power in a fixed effects model [94]. We discuss the implications of this result in subsequent sections.

We note that while equation (3.4) assumes $\mathbf{K}$ is the product of the centered and scaled genotype matrix [144], this relationship is also consistent with other genetic similarity metrics that yield a positive semi-definite $\mathbf{K}$. Other closely related metrics use the estimated rather than observed allele frequencies [150], adjust the similarity of an individual to itself to reduce sampling variation [215], use a Gower's centering to reduce sampling variance [83] or are proportional to these metrics [8]. Any of these similarity metrics can be used in the LMM or the principal components of the corresponding similarity matrix can be included as fixed effects.

3.3.2 Linear mixed model considers principal components' eigen-values

It is well established that the eigen-value of each principal component serves as a metric of biological relevance in relation to any underlying population structure [128, 144]. Thus a method for determining the relevance of a principal component to a given phenotype should consider both its eigen-value and its correlation with the phenotype [99]. Therefore, instead of considering only the principal components, $\mathbf{U}$, a more sophisticated model should consider the weighted principal components, $\mathbf{R} = \mathbf{US}$, since $\mathbf{S} = \text{diag}(\text{s})$ weights each principal component by its corresponding singular value (i.e. the square root of

its eigen-value). However, in the fixed effect model the estimate of the genetic effect, β, is invariant to the scale of the principal components due to the uniform prior implied in equation (3.2). Thus the fixed effect model assumes that each principal component has equal prior probability of being relevant to the phenotype. Alternatively, the LMM explicitly models the scale of the weighted principal components in equation (3.5). The LMM considers both the eigen-value and correlation with the phenotype when determining the relevance of each principal component to the phenotype. Moreover, the LMM's Gaussian prior on regression coefficients implies the biologically desirable property that a principal component with a larger eigen-value has a higher prior probability of being relevant to the phenotype [206].

3.3.3 Inference methods

Since modeling principal components as fixed or random effects share the underlying regression model, the differences in their ability to account for population structure and kinship [83, 87, 152, 207] can be attributed to the different inference methods and the number of principal components included. Yet the substantial theoretical and practical consequences of these differences have not been examined. With the goal of elucidating the statistical differences between modeling principal components as fixed versus random effects, we consider the theoretical properties of exact inference methods for the LMM [84, 108, 146, 230]. We note that our discussion also applies to approximate LMM methods since they approximate other aspects of the model [83, 182, 225].

In both fixed and random effects models, the parameter of interest for the hy-

pothesis test is the coefficient β corresponding to the effect of a single genetic marker, $\mathbf{x}_j$, so that the coefficients $\boldsymbol{\omega}$ or $\boldsymbol{\gamma}$ corresponding to the principal components are so-called nuisance parameters not of direct interest. The difference between the methods lies in how the statistical inference treats these nuisance parameters. The fixed effect model necessarily incorporates only $i \ll n$ principal components and maximizes the likelihood with respect to all coefficients so that the hypothesis test is conducted at the maximum likelihood estimates of the nuisance parameters. Thus the fixed effects model implies the likelihood

$$L_{fixed}(\beta, \mu, \boldsymbol{\omega}, \sigma_e^2 | \mathbf{y}) = \mathcal{N}(\mathbf{y} | \mu + \mathbf{x}_j \beta + \mathbf{U}_{1:i} \boldsymbol{\omega}, \sigma_e^2) \tag{3.6}$$

which has $i + 3$ free parameters to be estimated from the data. Therefore i degrees of freedom are used to correct for population structure.

Alternatively, the LMM includes all principal components in the model and integrates over the random effect with respect to its prior distribution. The likelihood can be stated in terms of the genetic similarity matrix,

$$L_{LMM}(\beta, \mu, \sigma_a^2, \sigma_e^2 | \mathbf{y}) \;\; = \;\; \int \mathcal{N}(\mathbf{y} | \mu + \mathbf{x}_j \beta + \boldsymbol{\alpha}, \sigma_e^2) \; \mathcal{N}(\boldsymbol{\alpha} | 0, \mathbf{K} \sigma_a^2) \; d\boldsymbol{\alpha} \tag{3.7}$$

[172] or equivalently in terms of the scaled principal components,

$$L_{LMM}(\beta, \mu, \sigma_a^2, \sigma_e^2 | \mathbf{y}) \;\; = \;\; \int \mathcal{N}(\mathbf{y} | \mu + \mathbf{x}_j \beta + \mathbf{R} \boldsymbol{\gamma}, \sigma_e^2) \; \mathcal{N}(\boldsymbol{\gamma} | 0, \sigma_a^2) \; d\boldsymbol{\gamma}$$

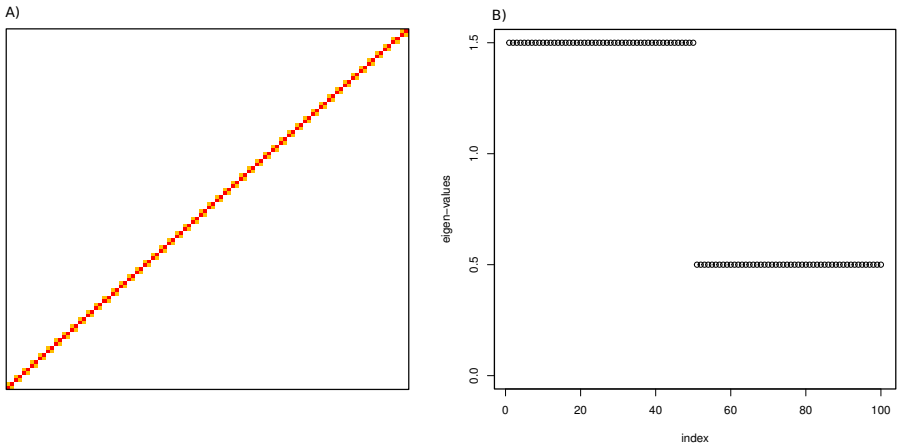
based on the equivalence between equations (3.3) and (3.5). While other equivalent forms of the likelihood are used for estimation in practice [84, 108], stating the likelihood in this way formalizes the Bayesian interpretation of the LMM where a Gaussian prior is placed on the regression coefficients of the principal components and the effect of population structure and kinship is integrated out. Due to the integration over nuisance parameters, the LMM is able to include all

principal components in the statistical model, yet estimate only 4 free parameters from the data.

3.3.4 Dimensionality of population structure versus kinship

Population structure and kinship are both confounding factors in GWAS since they produce covariance between individuals' phenotype values. Yet the dimensionality of these two processes are different. Population structure is a low dimensional process embedded in a high dimensional space so that a relatively small number of principal components represent the underlying population genetics [128, 144]. Therefore, a small number of principal components are adequate to account for population structure in GWAS datasets [150, 152]. Conversely, kinship is a high dimensional process since small sets of individuals are very closely related while being unrelated to the remaining individuals. Consider an idealized example of independent parent-offspring duos so that the coefficient of coancestry between parent and offspring is 0.5, and 0 between all other individuals. It follows directly that the corresponding coancestry matrix is block diagonal and the eigen-spectrum has a long tail so that all eigen-values are nonzero (Figure 3.1). In practice, the coefficient of coancestry is usually replaced by the observed genetic similarity so that the signature of kinship will not be as clean, and some trailing eigen-values will be due to stochastic noise. Nonetheless, kinship is a high-dimensional process that cannot be captured by a small number of principal components. Therefore, kinship must be modeled using a random effect as has long been done in the field of quantitative genetics [66, 117].

Figure 3.1: Genetic similarity matricies and their eigen-spectra. A) Block diagonal matrix of coefficients of coancestry for 50 parent-offspring duos. B) Eigen-spectrum of kinship matrix from (A).



3.3.5 Effective degrees of freedom of the linear mixed model

The degrees of freedom of a regression model is a widely used statistic that reflects the number of parameters estimated from the data. In a standard fixed effects model, the total degrees of freedom is a predetermined value and each regression coefficient uses 1 degree of freedom so that the total degrees of freedom equals the number of coefficients estimated. Moreover, the degrees of freedom of a fixed effects model must be substantially less than the sample size in order to maintain reasonable statistical power [94]. In standard GWAS analysis population structure is modeled as a low dimensional process [128, 144] so that the degrees of freedom devoted to the principal components is a small fraction of the sample size [150, 152]. Since the LMM is able to model both population structure and kinship by considering the full eigen-spectrum, it would therefore be useful to consider the degrees of freedom of the LMM. Due to the parameter-reduction property of the LMM resulting from the integration over nuisance parameters, we consider an analogous statistic, the "effective degrees

of freedom", df_e. We consider df_e in the context of the statistical properties of the LMM as well as its biological interpretation.

The formula for effective degrees of freedom is based on the fitted response values in the regression model. Ignoring fixed effects, since they each use 1 degree of freedom, the estimated trait values based on only the random effect are

$$\begin{aligned} \hat{\mathbf{y}} &= \mathbf{K}(\mathbf{K} + \mathbf{I}\delta)^{-1}\mathbf{y} \\ &= \mathbf{H}\mathbf{y} \end{aligned} \tag{3.8}$$

where $\delta = \frac{\sigma_e^2}{\sigma_a^2}$ and $\mathbf{H}$ is defined by construction [172]. This is recognizable as a "linear smoother", whereby the fitted response values are a linear function of the observed responses [64]. In such a case the effective degrees of freedom used by a linear smoother is defined as

$$df_e = tr(\mathbf{H}) \tag{3.9}$$

where $tr(.)$ gives the sum of the diagonal entries [64]. We note by way of comparison that a fixed effects model is also a linear smoother where $\hat{\mathbf{y}} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} = \mathbf{H}\mathbf{y}$ and that $tr(\mathbf{H})$ is equal to number of regression coefficients, thus satisfying our intuition about degrees of freedom. This metric of model complexity has seen wide adoption in the interpretation of penalized splines and nonparametric regression [64, 162]. The effective degrees of freedom for the LMM is more interpretable in the equivalent form

$$df_e = \sum_i \frac{s_i^2}{s_i^2 + \delta} \tag{3.10}$$

where s_i^2 is the i^{th} eigen-value of $\mathbf{K}$ forming the diagonal of $\mathbf{S}^2$ in equation (3.4) (see Appendix for derivation).

This form of the effective degrees of freedom facilitates an interpretation of the influence of each principal component that is composed of a marker-based element, s_i^2, and a phenotype based element, δ. It is apparent that the Gaussian prior in the LMM causes the influence of the i^{th} principal component to be a nonlinear function of the magnitude of its corresponding eigen-value, s_i^2. This formulation satisfies our intuition for df_e since the contribution of a single principal component is between 0 and 1 so that df_e, which is the sum of the contributions of all principal components, is naturally bounded between 0 and the number of principal components. Moreover, while s_i^2 has a local effect on the influence of each principal component separately and is independent of the phenotype, estimating δ adaptively learns the effective degrees of freedom based on the correlation of the principal components with the phenotype and has a global effect by influencing the contribution of all principal components. In addition, it follows that the effective degrees of each principal component decreases with its eigen-value.

Returning to the biological interpretation of the effective degrees of freedom, we note that LMM relates the genetic similarity between individuals to the heritability of the trait [215], as well as population structure and kinship [19, 83, 84, 150, 152]. Thus the LMM uses the estimated "pseudo-heritability" of the trait in the present set of individuals to determine how strongly to correct for population structure and kinship. This data-adaptive property reflects the ability of the LMM to learn df_e directly from the data. Moreover, the df_e statistic is composed of heritability, population structure and kinship so that the value of df_e reflects the "effective dimensionality" of the correction for confounding. Thus for a given heritability, a small df_e value indicates that only population

structure is relevant, while a large value indicates the high-dimensional kinship process is also relevant.

3.3.6 Low rank linear mixed model

To this point we have considered the standard LMM where the genetic similarity matrix is full rank and all principal components make a contribution to the phenotype. Yet the correction for confounding due to population structure and kinship should not necessarily be full rank. Including principal components that are not biologically relevant to the given phenotype can dilute the influence of relevant principal components and degrade the quality of the correction since the random effect is governed by a single global parameter, δ. Moreover, the biological relevance of a principal component to the phenotype depends both on its eigen-value and its correlation with the phenotype [99, 144]. Building on the algorithm of by Lippert et al. [108] for increasing computational efficiency for large datasets, we propose a novel method that uses the LRLMM framework and the effective degrees of freedom to identify principal components relevant to the current phenotype. In doing so we learn the rank of the correction for population structure and kinship.

In order to identify relevant principal components, we fit a LRLMM where the rank varies from 0, where we fit the standard linear model, to the sample size, where the full rank LMM is used. Principal components are added to the model sequentially and the log-likelihood and effective degrees of freedom are evaluated for each rank. The best model is then selected using the Bayesian Information Criterion (BIC) [168]. Since the order in which principal compo-

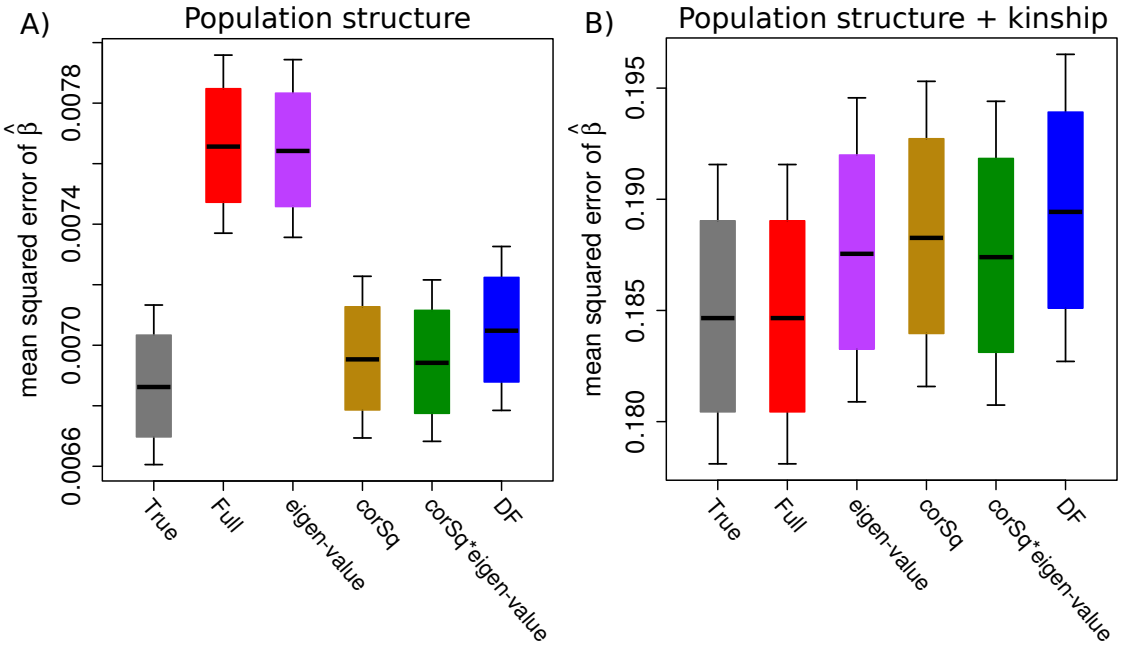
nents are added to the model affects the optimal rank, we consider different orderings of the principal components based on 1) eigen-value, 2) squared correlation between principal component and phenotype (corSq), 3) eigen-value multiplied by squared correlation between principal component and phenotype (corSq*eigen-value) [99], 4) degrees of freedom from fitting each principal component individually (DF).

3.4 Results

3.4.1 Simulations

Our low rank linear mixed model (LRLMM) is designed to empirically learn the optimal rank of the confounding effect of population structure and kinship. We assessed performance of the LRLMM by simulating continuous phenotypes using the normal model from equation (3.5) where a low rank model corresponds to setting most coefficients in $\boldsymbol{\gamma}$ to zero. We set $h^2 = \frac{\sigma_a^2}{\sigma_a^2+\sigma_e^2} = 0.40$, $\mu = \beta = 0$, drew nonzero coefficients from $\mathcal{N}(0, 100)$, and considered 2000 individuals with 10,000 replicate simulations. The genetic marker, $\mathbf{x}$, was generated based on an allele frequency drawn uniformly between 5% and 50%. Low dimensional population structure was modeled by randomly selecting 30 principal components from the first 90 to contribute to the phenotype. The high-dimensional process of kinship combined with population structure was modeled by having all principal components contribute to the phenotype. In both cases principal components are scaled by the square root of their eigen-value as in equation (3.5). The methods were assessed based on the accuracy of the estimates of genetic

Figure 3.2: Simulation results showing mean squared error of the genetic effect, $\hat{\beta}$, for linear mixed model (LMM) and low rank linear mixed model (LRLMM) for confounding due to A) population structure and B) population structure and kinship. Results are shown for the LRLMM using only the truly active principal components (True), full rank LMM (Full), and the LRLMM with four orderings of the principal components: eigen-value, squared correlation with phenotype (corSq), both eigen-value and squared correlation with phenotype (corSq*eigen-value) and degrees of freedom from fitting each principal component individually (DF). Bars indicate the estimated mean squared error of the genetic effect, $\hat{\beta}$, along with the 1%, 10%, 90% and 99% confidence interval.



effect expressed as the mean squared error (MSE) of $\hat{\beta}$ (Figure 3.2).

In the case of population structure the full-rank LMM had the largest MSE (Figure 3.2A). The LRLMM sorting principal components by correlation with the phenotype, correlation and eigen-value, or degrees of freedom had MSE close to that of the true model where only relevant principal components are used.

Sorting principal components by eigen-value also produced a large MSE since the eigen-value alone was not a good metric of the relevance of a principal component. In the presence of both population structure and kinship, the full-rank model is the true model so it yields the smallest MSE (Figure 3.2B). The LRLMM methods all give larger MSE values since they do not model the full eigen-spectrum of the genetic similarity matrix. In general, three LRLMM methods perform best when the confounding effect is low-dimensional, while the full-rank LMM performs best when when the confounding effect is also full-rank.

3.4.2 Data analysis

Our analysis of GWAS data from four populations and two phenotypes from the Multi-Ethnic Study of Atherosclerosis (MESA) [13] (Table 3.1) illustrates properties of the LMM and demonstrates the ability of the LRLMM to boost the strength of an association signal. Eigen-spectra of the genetic similarity matrices from four MESA populations as well as the matrix of coancestry coefficients based on the known pedigree from the Framingham Heart Study [32] illustrate the different dimensionality of population structure and kinship (Figure 3.3). It is apparent that population structure is low dimensional so the eigen-values decay very quickly in the MESA populations, while kinship from the Framingham pedigree shows a very long tail indicative of a high-dimensional process. In addition, the LMM relates the eigen-spectrum of the genetic similarity matrix to the phenotype and its heritability, and this relationship is reflected by the effective degrees of freedom for each principal component (Figure 3.4). Thus the effective degrees of freedom, normalized by the sample size, used by the LMM for height is substantially larger than for HDL cholesterol (Figure 3.6A), since

height is known to have a larger heritability [95, 204]. Moreover, the fact that the effective degrees of freedom is a substantial fraction of the sample size indicates that the LMM models the high-dimensional confounding effect of kinship. We note that the heterogeneity among populations can be attributed either to differential population structure or kinship, or to stochastic effects. Finally, we note that the LMM was fit by maximum likelihood here, but estimation by REML has little effect (Figure 3.5).

Table 3.1: Sample size for each population and phenotype from the Multi-Ethnic Study of Atherosclerosis (MESA) dataset.

	Asian	Hispanic	European	African American	combined
HDL cholesterol	772	1436	2481	1584	6273
height	775	2104	2522	2528	7929

Applying the LRLMM sorting by degrees of freedom from fitting each principal component individually (LRLMM-DF) selects effective degrees of freedom that are substantially smaller than for the full-rank model and the effective degrees of freedom is generally larger for height than for HDL cholesterol (Figure 3.6B). Morever, the width of the 95% confidence interval is also substantially smaller. Applying the LRLMM-DF for association testing for HDL cholesterol in Europeans substantially boosts the signal from markers on chr8 between positions 19,852,309 and 19,869,675 compared to a standard linear model (Plink [154]) and three versions of the LMM (EMMAX [83], GEMMA [230], fastlmm [108]) (Figure 3.7, Figure A.23). The boost in the association signal is more interpretable in a zoom-in manhattan plot illustrating that the LRLMM-DF method produces many more p-values that exceed the Bonferroni cutoff (Figure 3.8). This region has previously been associated with HDL cholesterol [93, 203], so LRLMM-DF is able to strengthen the signal of a replicated association.

Figure 3.3: Comparison of eigen-spectra due to population structure and kinship. The eigen-spectrum based on the known pedigree from 3063 individuals from the Framingham Heart Study reflects kinship, while the eigen-spectrum for four populations from the Multi-Ethnic Study of Atherosclerosis (MESA) reflects both population structure and kinship. Eigen-values for each dataset are normalized by the maximum eigen-value so that each spectrum has a maximum of 1.

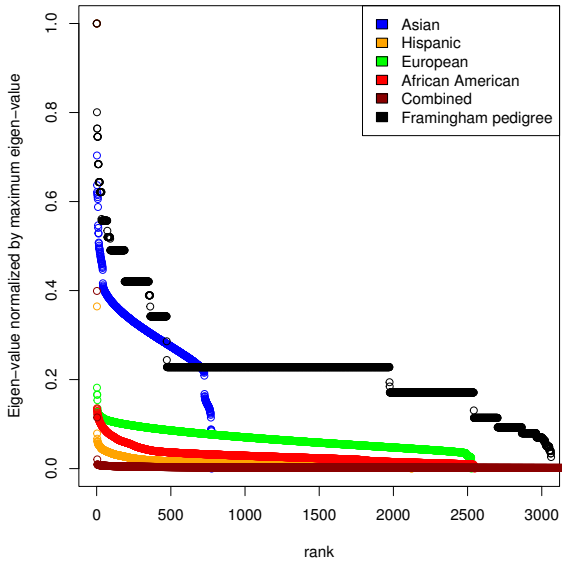


Figure 3.4: Effective degrees of freedom for each principal component based on a linear mixed model (LMM) analysis of HDL cholesterol for four populations from the Multi-Ethnic Study of Atherosclerosis (MESA) dataset. Total effective degrees of freedom for each population are shown in the legend.

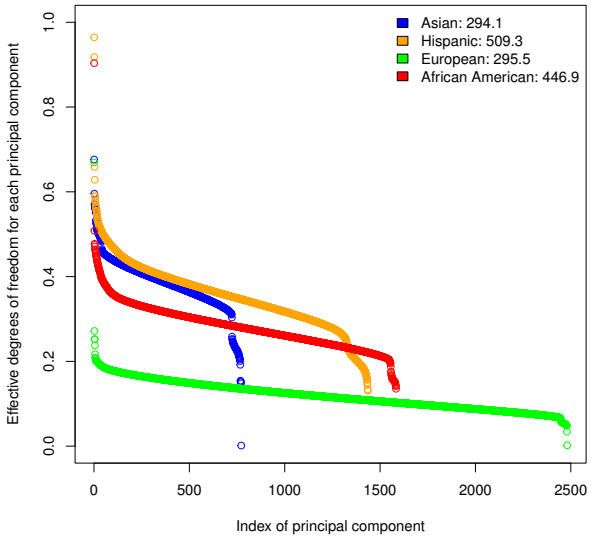


Figure 3.5: Fraction of available degrees of freedom used by the linear mixed model (LMM) to account for population structure and kinship estimated using restricted maximum likelihood (REML). Effective degrees of freedom normalized by sample size are show for six phenotypes and four populations from the Multi-Ethnic Study of Atherosclerosis (MESA) plus the combined dataset. Error bars indicate 95% confidence intervals based on the log-likelihood surface.

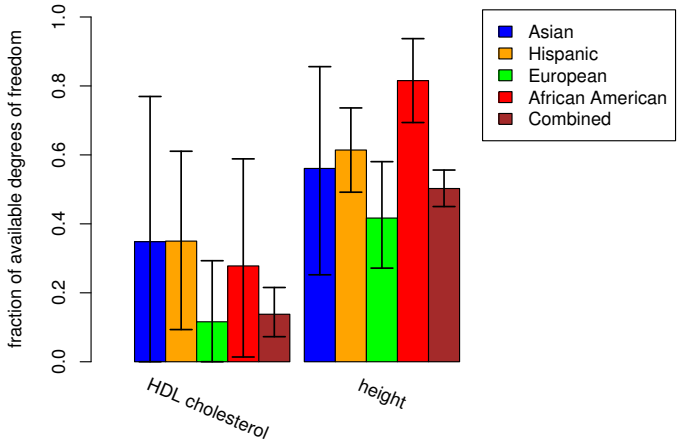


Figure 3.6: Fraction of available degrees of freedom used to account for population structure and kinship by A) the linear mixed model (LMM) and B) the low rank linear mixed model (LRLMM) sorting by degrees of freedom of each principal component fit individually (LRLMM-DF). Effective degrees of freedom normalized by sample size are shown for two phenotypes and four populations from the Multi-Ethnic Study of Atherosclerosis (MESA) plus the combined dataset. Error bars indicate 95% confidence intervals based on the log-likelihood surface. Note the large difference in the scales between (A) and (B).

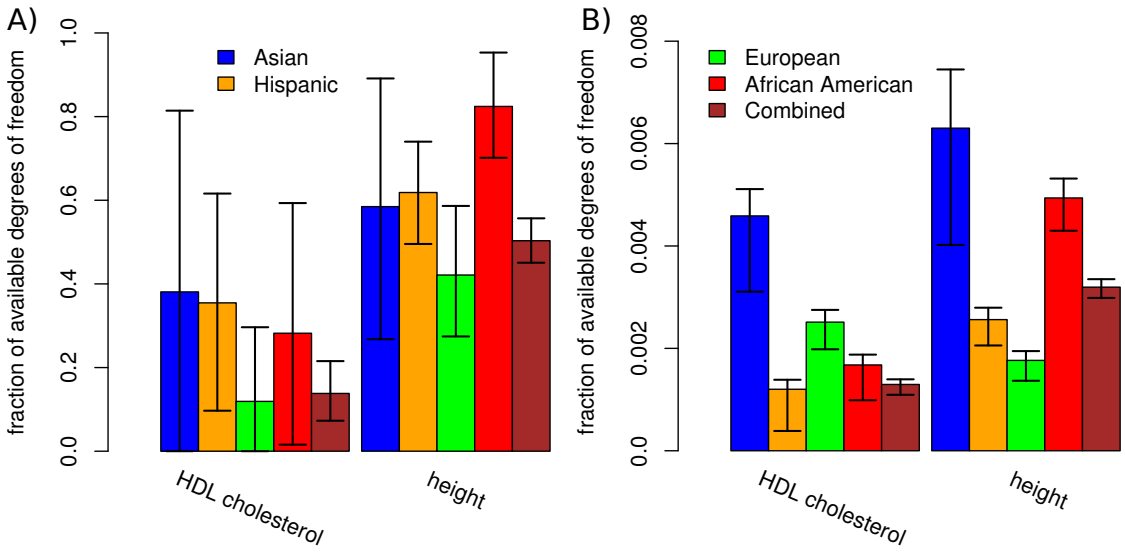


Figure 3.7: Quantile-quantile plot for association tests for HDL cholesterol in Europeans. Results are shown from a standard linear model (Plink [154]), 3 versions of the linear mixed model (EMMAX [83], GEMMA [230], fastlmm [108]), and the low rank linear mixed model with 4 orderings of the principal components. We note that LRLMM using eigen-value and corSq*eigen-value orderings selected no principal components correction and thus give the same p-values as Plink. λ_{GC} indicates the genomic control value [34].

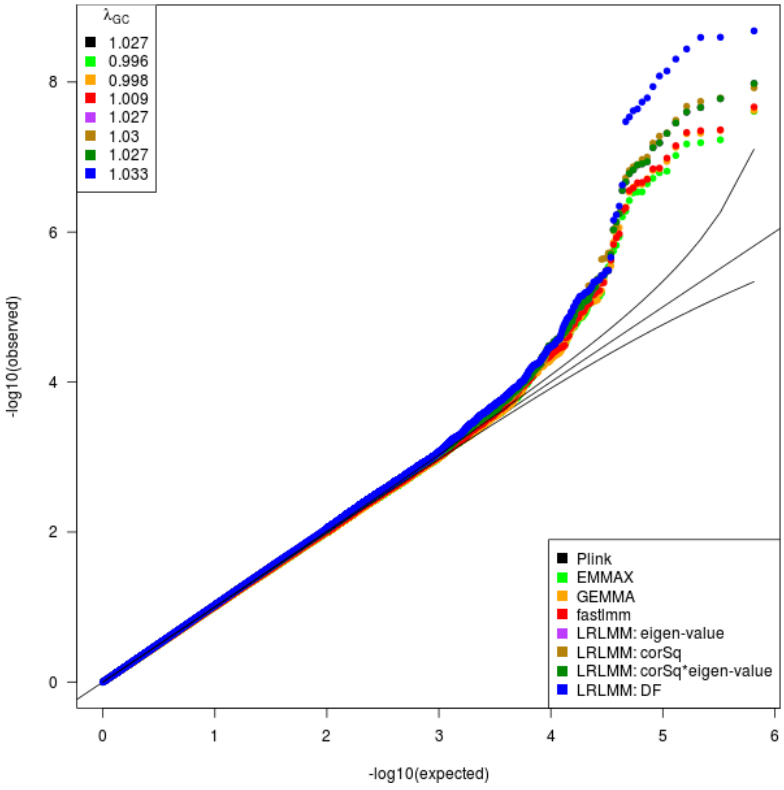
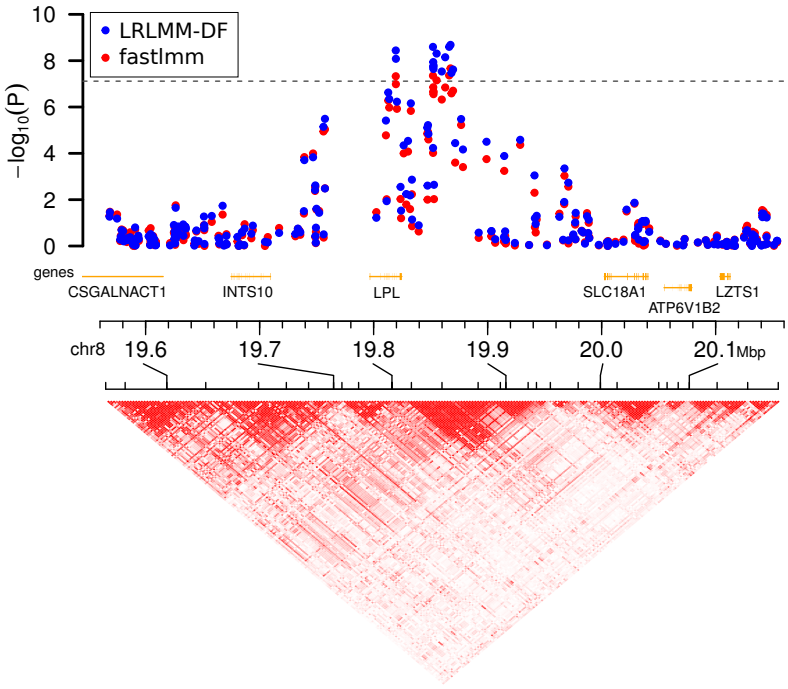


Figure 3.8: Manhattan plot of chromosome 8 showing 19.6 Mbp to 20.1 Mbp where the low-rank linear mixed model (LRLMM) ordering principal components by degrees of freedom based on the fit of LRLMM with each principal component individually (LRLMM-DF). P-values from fastlmm [108] are shown for comparison. Dashed line indicates Bonferroni correction of 5% for 650,000 markers. Linkage disequilibrium is shown in terms of D′. We note that other methods are omitted for the sake of clarity.



3.5 Discussion

The linear mixed model (LMM) has become a standard method to account for the confounding effects of population structure and kinship in GWAS datasets. [83, 152]. Our theoretical and empirical analysis illustrates the properties of the LMM and formalizes a biological interpretation of the model. We introduced the effective degrees of freedom in order to interpret the strength of the correction for population structure and kinship. A fixed effects model can include relatively few principal components, yet the LMM models the entire eigenspectrum of the genetic similarity matrix. Thus while it is generally suggested that the degrees of freedom in a regression model be on the order of the square-root of the sample size in order to maintain reasonable statistical power [94], the effective degrees of freedom of the full-rank LMM routinely exceeded 40% of the sample size in our analysis and reached up to 80%. The effect of using such high effective degrees of freedom on the statistical test of association remains an open question. Moreover, wide confidence intervals for the effective degrees of freedom indicate that there is a high degree of uncertainly about the strength of the correction for population structure and kinship. Alternatively, the confidence intervals for the LRLMM are substantially smaller and are thus less influenced by stochastic effects. These results indicate that a high-dimensional correction for confounding may benefit from a fully Bayesian treatment of the linear mixed model as it would integrate over the uncertainty of the strength of the correction. Yet the LRLMM would likely not benefit as much since it produces a low-dimensional fit to the data.

The ability of our low rank linear mixed model (LRLMM) to boost the signal of

a known association for HDL cholesterol in Europeans indicates that the LMM can overfit the data so that the random effect absorbs too much of the phenotype variance. If the true model is low rank, then the LRLMM will have greater power than the LMM. Alternatively, if the true model is high-dimensional then the full-rank LMM is more appropriate. Since there is no principled way to determine the true dimensionality, our novel LRLMM provides an alternative test of association that can boost the strength of an association or identify additional associations if it is a better fit to the data. As there is currently no objective way to determine which method is more appropriate for a given dataset, we recommend running both the full and low rank methods and examining the resulting quantile-quantile plots and top associations.

With the growing interest in testing associations of rare variants, new problems of population structure are arising due to the more recent origin and more localized distribution of rare compared to common variants [85, 136, 184]. Moreover, a recent simulation study demonstrated that including principal components as covariates or using the LMM were not effective at controlling for population stratification due to rare variants [123]. While addressing this challenge will require extensive methodology development and empirical investigations, the framework discussed here suggests important issues to consider in order to apply appropriate corrections for population structure and kinship in the next-generation of GWAS.

CHAPTER 4

MODELING THE POLYGENIC ARCHITECTURE OF COMPLEX TRAITS IN THE PRESENCE OF KINSHIP AND POPULATION STRUCTURE: A UNIFIED STATISTICAL FRAMEWORK

Sophisticated statistical methods that model the biological complexities of genome-wide association studies (GWAS) can increase the power to detect weak associations while controlling the false discovery rate. Among the most successful methodological developments has been the use of linear mixed models to correct for kinship and population structure [83, 84, 92, 108, 109, 146, 156, 170, 182, 230]. Independently, multiple-locus methods that explicitly model the polygenic architecture of complex traits have the potential to increase power to detect weak associations [9, 39, 57, 65, 72, 103, 114, 192, 209, 214, 227], and we have previously demonstrated the power of this approach (Chapter 2). Here we combine these complementary statistical methods into a unified statistical framework. This penalized linear mixed model is scalable to large datasets and has direct applications to response prediction and feature selection in the presence of genetic confounding.

The linear mixed model has previously been combined with feature select methods using a LASSO penalty [156] or stepwise regression [170]. Here we propose statistical, algorithmic and computational developments in order to optimize the statistical performance and facilitate scaling to large datasets. We develop novel approaches to tuning nonconvex penalties and determining the optimal stopping point in regularization path. Leveraging recent work on assessing significance of selected features, we produce a well-principled and scalable statistical method applicable to feature selection, hypothesis testing and prediction

in many contexts. We are currently developing a user-friendly R package based on these developments.

4.1 Methods

4.1.1 Linear mixed model

Consider the linear mixed model widely used in GWAS analysis [83, 84, 92, 108, 109, 146, 170, 230] and derive parameter updates to reach a local model of the likelihood. The log-likelihood can be expressed as

$$\ell(\boldsymbol{\beta}, \sigma_e^2, \sigma_a^2) = \log \mathcal{N}\left[\mathbf{y}|\mathbf{X}\boldsymbol{\beta}, \mathbf{K}\sigma_a^2 + \mathbf{I}\sigma_e^2\right] \tag{4.1}$$

$$= \log \mathcal{N}\left[\mathbf{y}|\mathbf{X}\boldsymbol{\beta}, \sigma_a^2(\mathbf{K} + \mathbf{I}\delta)\right] \qquad \text{where } \delta = \frac{\sigma_e^2}{\sigma_a^2} \tag{4.2}$$

$$= -\frac{n}{2}\log(2\pi) - \frac{1}{2}\log|\sigma_a^2(\mathbf{K} + \mathbf{I}\delta)| - \frac{1}{2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T(\sigma_a^2(\mathbf{K} + \mathbf{I}\delta))^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \tag{4.3}$$

$$= -\frac{n}{2}\log(2\pi\sigma_a^2) - \frac{1}{2}\log|\mathbf{K} + \mathbf{I}\delta| - \frac{1}{2\sigma_a^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{K} + \mathbf{I}\delta)^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \tag{4.4}$$

Letting the eigen-decomposition of $\mathbf{K} = \mathbf{U}\mathbf{S}\mathbf{U}^T$,

$$\mathbf{K} + \mathbf{I}\delta = \mathbf{U}\mathbf{S}\mathbf{U}^T + \mathbf{U}\mathbf{U}^T\delta \tag{4.5}$$

$$= \mathbf{U}(\mathbf{S} + \mathbf{I}\delta)\mathbf{U}^T \tag{4.6}$$

Plugging (4.6) into the log-likelihood (4.4), the log-likelihood becomes

$$= -\frac{n}{2}\log(2\pi\sigma_a^2) - \frac{1}{2}\log|\mathbf{U}(\mathbf{S} + \mathbf{I}\delta)\mathbf{U}^T| - \frac{1}{2\sigma_a^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{U}(\mathbf{S} + \mathbf{I}\delta)\mathbf{U}^T)^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \tag{4.7}$$

$$= -\frac{n}{2}\log(2\pi\sigma_a^2) - \frac{1}{2}\log|\mathbf{S} + \mathbf{I}\delta| - \frac{1}{2\sigma_a^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T\mathbf{U}(\mathbf{S} + \mathbf{I}\delta)^{-1}\mathbf{U}^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \tag{4.8}$$

$$= -\frac{n}{2}\log(2\pi\sigma_a^2) - \frac{1}{2}\log|\mathbf{S} + \mathbf{I}\delta| - \frac{1}{2\sigma_a^2}([\mathbf{U}^T\mathbf{y}] - [\mathbf{U}^T\mathbf{X}]\boldsymbol{\beta})^T(\mathbf{S} + \mathbf{I}\delta)^{-1}([\mathbf{U}^T\mathbf{y}] - [\mathbf{U}^T\mathbf{X}]\boldsymbol{\beta}) \tag{4.9}$$

Letting $\tilde{\mathbf{y}} = \mathbf{U}^T\mathbf{y}$ and $\tilde{\mathbf{X}} = \mathbf{U}^T\mathbf{X}$ in order to simplify notation,

$$= -\frac{1}{2}\left[n\log(2\pi\sigma_a^2) + \log|\mathbf{S}+\mathbf{I}\delta| + \frac{1}{\sigma_a^2}(\tilde{\mathbf{y}}-\tilde{\mathbf{X}}\boldsymbol{\beta})^T(\mathbf{S}+\mathbf{I}\delta)^{-1}(\tilde{\mathbf{y}}-\tilde{\mathbf{X}}\boldsymbol{\beta})\right] \quad (4.10)$$

Since the covariance matrix is now diagonal, the log-likelihood

$$\ell(\boldsymbol{\beta},\sigma_e^2,\sigma_a^2) = -\frac{1}{2}\left[n\log(2\pi\sigma_a^2) + \sum_{i=1}^{n}\log(\mathbf{S}_{i,i}+\delta) + \frac{1}{\sigma_a^2}\sum_{i=1}^{n}\frac{(\tilde{\mathbf{y}}_i-\tilde{\mathbf{X}}_i\boldsymbol{\beta})^2}{\mathbf{S}_{i,i}+\delta}\right] \quad (4.11)$$

can be evaluate in time linear with the sample size.

Estimation

Taking derivatives of (4.11) gives closed form updates of σ_a^2 and $\boldsymbol{\beta}$:

$$\frac{\partial\ell}{\partial\sigma_a^2} = -\frac{1}{2}\left(\frac{n}{\sigma_a^2} - \frac{1}{\sigma_a^4}\sum_{i=1}^{n}\frac{(\tilde{\mathbf{y}}_i-\tilde{\mathbf{X}}_i\boldsymbol{\beta})^2}{\mathbf{S}_{i,i}+\delta}\right) = 0 \quad (4.12)$$

$$\hat{\sigma}_a^2 = \frac{1}{n}\sum_{i=1}^{n}\frac{(\tilde{\mathbf{y}}_i-\tilde{\mathbf{X}}_i\boldsymbol{\beta})^2}{\mathbf{S}_{i,i}+\delta} \quad (4.13)$$

$$\frac{\partial\ell}{\partial\boldsymbol{\beta}} = \frac{1}{\sigma_a^2}(\tilde{\mathbf{y}}-\tilde{\mathbf{X}}\boldsymbol{\beta})^T(\mathbf{S}+\mathbf{I}\delta)^{-1}\tilde{\mathbf{X}} = 0 \quad (4.14)$$

$$= \tilde{\mathbf{y}}^T(\mathbf{S}+\mathbf{I}\delta)^{-1}\tilde{\mathbf{X}} - \boldsymbol{\beta}^T\tilde{\mathbf{X}}^T(\mathbf{S}+\mathbf{I}\delta)^{-1}\tilde{\mathbf{X}} \quad (4.15)$$

$$\hat{\boldsymbol{\beta}} = \left[\tilde{\mathbf{X}}^T(\mathbf{S}+\mathbf{I}\delta)^{-1}\tilde{\mathbf{X}}\right]^{-1}\tilde{\mathbf{X}}^T(\mathbf{S}+\mathbf{I}\delta)^{-1}\tilde{\mathbf{y}} \quad (4.16)$$

$$= \left[\sum_{i=1}^{n}\frac{1}{\mathbf{S}_{i,i}+\delta}\tilde{\mathbf{X}}_i^T\tilde{\mathbf{X}}_i\right]^{-1}\left[\sum_{i=1}^{n}\frac{1}{\mathbf{S}_{i,i}+\delta}\tilde{\mathbf{X}}_i^T\tilde{\mathbf{y}}_i\right] \quad (4.17)$$

Since updates of σ_a^2 and $\boldsymbol{\beta}$ have closed forms, the update of the δ becomes a one dimensional optimization problem by plugging in values of the other parameters. The log-likelihood can thus be expressed as

$$\ell(\delta) = -\frac{1}{2}\left[n\log(2\pi\hat{\sigma}_a^2) + \sum_{i=1}^{n}\log(\mathbf{S}_{i,i}+\delta) + n\right] \quad (4.18)$$

which is a simple function of δ. Since the log-likelihood surface is not convex with respect to δ we apply a one dimensional grid search with a quasi-Newton optimization method to identify modes of the surface between each

pair of points on the grid.

The derivation of this part of the model follows Lippert et al. [108] and we omit the derivation of the low rank mixed model for brevity.

4.1.2 Penalized linear mixed model

Combining the linear mixed model with a penalty on the regression coefficients, $\boldsymbol{\beta}$, involves including a penalty term in the log-likelihood so that the penalized likelihood has the form

$$\begin{aligned} \ell(\boldsymbol{\beta}, \sigma_a^2, \delta) &= \log \mathcal{N}\left[\mathbf{y}|\mathbf{X}\boldsymbol{\beta}, \sigma_a^2(\mathbf{K} + \mathbf{I}\delta)\right] - p_\lambda(\boldsymbol{\beta}) && (4.19) \\ &= -\frac{1}{2}\left[n\log(2\pi\sigma_a^2) + \sum_{i=1}^{n}\log(\mathbf{S}_{i,i} + \delta) + \frac{1}{\sigma_a^2}(\tilde{\mathbf{y}} - \tilde{\mathbf{X}}\boldsymbol{\beta})^T(\mathbf{S} + \mathbf{I}\delta)^{-1}(\tilde{\mathbf{y}} - \tilde{\mathbf{X}}\boldsymbol{\beta})\right] \\ &\quad - \sum_j p(\beta_j) && (4.20) \end{aligned}$$

for an arbitrary penalty $p(.)$.

Estimation

Maximizing the log-likelihood involves iteratively updating each parameter. Since the updates of σ_a^2 and δ only depend on $\boldsymbol{\beta}$ through the current residuals $\tilde{\mathbf{r}} = \tilde{\mathbf{y}} - \tilde{\mathbf{X}}\hat{\boldsymbol{\beta}}$, the updates of these parameters remains the same as for the unpenalized model. The regression coefficients, $\boldsymbol{\beta}$, are updated using a coordinate-wise update approached that is scalable to high-dimensional penalized regression [47, 48].

Considering only terms in the log-likelihood that are relevant to the updates

of $\boldsymbol{\beta}$, the log-likelihood is recognizable as a reweighted penalized regression system [48]. Letting the weights be $\mathbf{W} = (\mathbf{S} + \mathbf{I}\delta)^{-1}$, separating out the j^{th} regression coefficient, and using the fact that

$$\tilde{\mathbf{y}} - \sum_{k \neq j} \tilde{\mathbf{x}}_k \beta_k - \tilde{\mathbf{x}}_j \beta_j \;\; = \;\; \tilde{\mathbf{r}} + \tilde{\mathbf{x}}_j \beta_j \tag{4.21}$$

the log-likelihood can be expressed as

$$\begin{aligned}
\ell(\boldsymbol{\beta}, \sigma_a^2, \delta) \;\; &\propto \;\; -\frac{1}{2\sigma_a^2}(\tilde{\mathbf{y}} - \tilde{\mathbf{X}}\boldsymbol{\beta})^T(\mathbf{S} + \mathbf{I}\delta)^{-1}(\tilde{\mathbf{y}} - \tilde{\mathbf{X}}\boldsymbol{\beta}) - \sum_j p(\beta_j) && (4.22)\\
&= \;\; -\frac{1}{2\sigma_a^2}(\tilde{\mathbf{y}} - \sum_{k \neq j} \tilde{\mathbf{x}}_k \beta_k - \tilde{\mathbf{x}}_j \beta_j)^T \mathbf{W} (\tilde{\mathbf{y}} - \sum_{k \neq j} \tilde{\mathbf{x}}_k \beta_k - \tilde{\mathbf{x}}_j \beta_j) - \sum_j p(\beta_j) && (4.23)\\
&= \;\; -\frac{1}{2\sigma_a^2}(\tilde{\mathbf{r}} + \tilde{\mathbf{x}}_j \beta_j)^T \mathbf{W} (\tilde{\mathbf{r}} + \tilde{\mathbf{x}}_j \beta_j) - \sum_j p(\beta_j) && (4.24)
\end{aligned}$$

It follows that the first and second derivatives are

$$\begin{aligned}
\frac{\partial \ell}{\partial \beta_j} \;\; &= \;\; \frac{1}{\sigma_a^2}\left[(\tilde{\mathbf{r}} + \tilde{\mathbf{x}}_j \beta_j)^T \mathbf{W} \tilde{\mathbf{x}}_j - \tilde{\mathbf{x}}_j^T \mathbf{W} \tilde{\mathbf{x}}_j \beta_j\right] - \sum_j p'(\beta_j) && (4.25)\\
&= \;\; \frac{1}{\sigma_a^2}\left[\tilde{\mathbf{r}}^T \mathbf{W} \tilde{\mathbf{x}}_j + \tilde{\mathbf{x}}_j^T \mathbf{W} \tilde{\mathbf{x}}_j \beta_j - \tilde{\mathbf{x}}_j^T \mathbf{W} \tilde{\mathbf{x}}_j \beta_j\right] - \sum_j p'(\beta_j) && (4.26)\\
&= \;\; \frac{1}{\sigma_a^2} \tilde{\mathbf{r}}^T \mathbf{W} \tilde{\mathbf{x}}_j - \sum_j p'(\beta_j) && (4.27)\\
\frac{\partial \ell}{\partial \beta_j^2} \;\; &= \;\; -\frac{1}{\sigma_a^2} \tilde{\mathbf{x}}_j^T \mathbf{W} \tilde{\mathbf{x}}_j - \sum_j p''(\beta_j) && (4.28)
\end{aligned}$$

so that the standard Newton-Raphson update with the LASSO model is

$$\beta_j^{(s+1)} \;\; = \;\; \beta_j^{(s)} - \frac{\ell'(\beta_j^{(s)})}{\ell''(\beta_j^{(s)})}. \tag{4.29}$$

Note that this derivation and those below omit the complexities raised by the fact that $p(\beta)$ is non-differentiable at the origin. We follow Wu and Lange [210] on this issue, but omit the details for brevity.

Estimation: LASSO

For the LASSO penalty [187] the form of the update is

$$\begin{aligned}
\beta_j^{(s+1)} &= \beta_j^{(s)} + \frac{\frac{1}{\sigma_a^2}\tilde{\mathbf{r}}^T\mathbf{W}\tilde{\mathbf{x}}_j - \lambda sign(\beta_j)}{\frac{1}{\sigma_a^2}\tilde{\mathbf{x}}_j^T\mathbf{W}\tilde{\mathbf{x}}_j} && (4.30)\\
&= \frac{\frac{1}{\sigma_a^2}\beta_j^{(s)}\tilde{\mathbf{x}}_j^T\mathbf{W}\tilde{\mathbf{x}}_j + \frac{1}{\sigma_a^2}\tilde{\mathbf{r}}^T\mathbf{W}\tilde{\mathbf{x}}_j - \lambda sign(\beta_j)}{\frac{1}{\sigma_a^2}\tilde{\mathbf{x}}_j^T\mathbf{W}\tilde{\mathbf{x}}_j} && (4.31)\\
&= \frac{\beta_j^{(s)}\tilde{\mathbf{x}}_j^T\mathbf{W}\tilde{\mathbf{x}}_j + \tilde{\mathbf{r}}^T\mathbf{W}\tilde{\mathbf{x}}_j - \sigma_a^2\lambda sign(\beta_j)}{\tilde{\mathbf{x}}_j^T\mathbf{W}\tilde{\mathbf{x}}_j} && (4.32)
\end{aligned}$$

Letting $z = \beta_j^{(s)}\tilde{\mathbf{x}}_j^T\mathbf{W}\tilde{\mathbf{x}}_j + \tilde{\mathbf{r}}^T\mathbf{W}\tilde{\mathbf{x}}_j$, the LASSO update is

$$\beta_j^{(s+1)} = \frac{S(z, \sigma_a^2\lambda)}{\tilde{\mathbf{x}}_j^T\mathbf{W}\tilde{\mathbf{x}}_j} \tag{4.33}$$

where

$$S(d,\lambda) = \begin{cases} d-\lambda & d > 0 \text{ and } \lambda < |d| \\ d+\lambda & d < 0 \text{ and } \lambda < |d| \\ 0 & \lambda \geq |d| \end{cases} \tag{4.34}$$

is the soft-thresholding function [47, 48].

Estimation: MCP

If an MCP penalty [221] is used, then the log-likelihood becomes

$$\ell(\boldsymbol{\beta}, \sigma_e^2, \sigma_a^2) = \log \mathcal{N}\left[\mathbf{y}|\mathbf{X}\boldsymbol{\beta}, \mathbf{K}\sigma_a^2 + \mathbf{I}\sigma_e^2\right] - \sum_j p_{\lambda,a}(\beta_j) \tag{4.35}$$

where

$$p_{\lambda,a}(\beta) = \begin{cases} \lambda\beta - \frac{\beta^2}{2a} & \beta \leq a\lambda \\ \frac{1}{2}a\lambda^2 & \beta > a\lambda \end{cases} \tag{4.36}$$

It follows the directly from above derivations and the derivation of penalized likelihood models for MCP in Chapter 2 that coordinate-wise update for $\beta \leq a\lambda$ is

$$\begin{aligned}
\beta_j^{(s+1)} &= \beta_j^{(s)} + \frac{\frac{1}{\sigma_a^2}\mathbf{r}^T\mathbf{W}\mathbf{x}_j - \lambda + \frac{\beta_j^{(s)}}{a}}{\frac{1}{\sigma_a^2}\mathbf{x}_j^T\mathbf{W}\mathbf{x}_j - \frac{1}{a}} && (4.37)\\
&= \frac{\beta_j^{(s)}(\frac{1}{\sigma_a^2}\mathbf{x}_j^T\mathbf{W}\mathbf{x}_j - \frac{1}{a}) + \frac{1}{\sigma_a^2}\mathbf{r}^T\mathbf{W}\mathbf{x}_j - \lambda + \frac{\beta_j^{(s)}}{a}}{\frac{1}{\sigma_a^2}\mathbf{x}_j^T\mathbf{W}\mathbf{x}_j - \frac{1}{a}} && (4.38)\\
&= \frac{\frac{1}{\sigma_a^2}\beta_j^{(s)}\mathbf{x}_j^T\mathbf{W}\mathbf{x}_j + \frac{1}{\sigma_a^2}\mathbf{r}^T\mathbf{W}\mathbf{x}_j - \lambda}{\frac{1}{\sigma_a^2}\mathbf{x}_j^T\mathbf{W}\mathbf{x}_j - \frac{1}{a}} && (4.39)\\
&= \frac{\beta_j^{(s)}\mathbf{x}_j^T\mathbf{W}\mathbf{x}_j + \mathbf{r}^T\mathbf{W}\mathbf{x}_j - \sigma_a^2\lambda}{\mathbf{x}_j^T\mathbf{W}\mathbf{x}_j - \frac{\sigma_a^2}{a}} && (4.40)\\
&= \frac{S(z, \sigma_a^2\lambda)}{\mathbf{x}_j^T\mathbf{W}\mathbf{x}_j - \frac{\sigma_a^2}{a}} && (4.41)
\end{aligned}$$

4.1.3 A hypothesis test of selected features

Hypothesis testing is often the ultimate goal of regression modeling. Yet feature selection methods do not typically return a p-value or other confidence metric that reflects the significance of each selected feature. Recent work has produced a number of subsampling methods that perform feature selection on a subset of the data and either perform hypothesis testing on a complementary subset [202], perform this process on many subsets to produce a distribution of p-values [131], or return the fraction of times that each feature is selected [130]. Yet these methods are computationally expensive and are problematic when considering highly correlated features, such as in GWAS datasets [2].

In recent work, Lockhart et al. [112] has developed a very simple hypothesis test for the addition of each feature as the penalty is gradually relaxed. The simplic-

ity of the test statistic makes it directly applicable to penalized feature selection models using LASSO, MCP or another penalty. Moreover, the test should be approximately correct for the penalized linear mixed models described here.

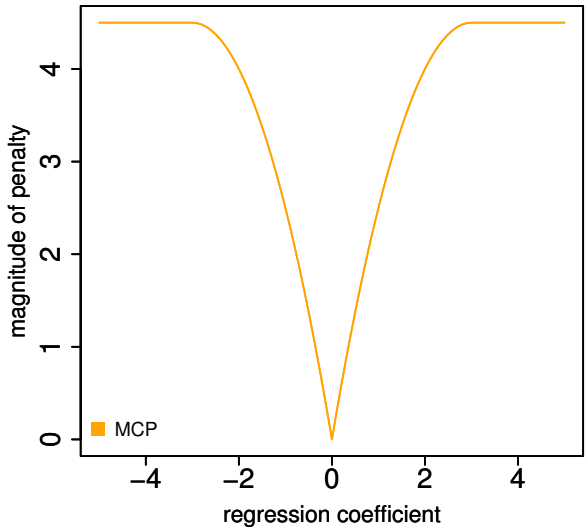
4.1.4 Determining optimal tuning parameter from regularization path

Selecting the optimal value of the tuning parameter governing the slope at the origin as been widely studied [26, 48, 198, 199, 211, 224], yet still remains an issue in data analysis. Based on the work of Lockhart et al. [112], a hypothesis test produces an interpretable confidence metric for each value of the tuning parameter and can serve as the criterion for determining the stopping point in the regularization path. We note that this approach naturally incorporates both the sample size and the number of features into the stopping criteria.

4.1.5 Selecting second tuning parameter for MCP

MCP has two tuning parameters, λ, which determines the derivative infinitesimally close to the origin, and a so that the penalty becomes constant at $a\lambda$ (Equation 4.36, Figure 4.1). While tuning the slope near the origin has been well studied, selecting a remains an open question. Here we propose a novel strategy based on fixing the value of $a\lambda$. We note that the MCP no longer shrinks coefficient estimates that surpass this value, since the derivative is zero. Essentially, this is akin to saying that we have high confidence that all coefficients greater than $a\lambda$ are truly nonzero. This interpretation motivates the question:

Figure 4.1: The MCP penalty



For a given dataset, how large must a coefficient estimate be before we have high confidence that it is truly nonzero? We can thus set $a\lambda$ to the value that answers this query and set a based on the value of λ. Formalizing this intuitive motivation based on statistical theory, we can determine the value $a\lambda$ by referring to standard power calculations for the linear model with Gaussian error. Based on the dataset given and the desired confidence level this theory yields a simple form the value of $a\lambda$.

Consider a Gaussian linear model of the form

$$\mathbf{y} = \mu + \sum_{j=1}^{m} \mathbf{x}_j \beta_j + \boldsymbol{\varepsilon} \tag{4.42}$$

$$\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \sigma^2) \tag{4.43}$$

with n samples. Using standard results from statistical theory [23], we can assess the power of an F-test of the j^{th} coefficient, β_j, to reject the null hypothesis of $\beta_j = 0$ based on n, σ^2, $var(\mathbf{x}_j)$, α (the type I error rate), and the value of β_j

under the alternative. In general, the power can be expressed as

$$\text{power} \quad = \quad p(F(1, n-m-1, \gamma) \geq t) \tag{4.44}$$

where $F(1, n-m-1, \gamma)$ is the cumulative distribution function of the F-distribution, $t = F_{1-\alpha,1,n-m-1}$ is the critical value based on the $1-\alpha$ quantile of the F-distribution, γ is the test statistic defined as

$$\gamma = \frac{n\beta_j^2(1 - R^2_{\mathbf{x}_j|\mathbf{x}_{-j}})var(\mathbf{x}_j)}{\sigma^2} \tag{4.45}$$

and $R^2_{\mathbf{x}_j|\mathbf{x}_{-j}}$ denotes a conditional correlation between $\mathbf{x}_j$ and all other variables.

For simplicity, we can deal with a univariate regression (with a mean term) so that the test statistic becomes

$$\gamma = \frac{n\beta_j^2 var(\mathbf{x}_j)}{\sigma^2} \tag{4.46}$$

and power becomes

$$\text{power} \quad = \quad p(F(1, n-2, \gamma) \geq F_{1-\alpha,1,n-2}). \tag{4.47}$$

Based on this theory, the value of $a\lambda$ can be expressed as a function of properties of the data n and $var(\mathbf{x}_j)$; an estimate of the residual variance, σ^2, which can be informed by the estimated heritability of the trait; the desired confidence level, expressed as the desired type I error, which can be informed by the multiple testing burden; and the desired power. We note that this approach depends on the MCP penalty reaching a constant value for a finite coefficient value. Since the LOG and NEG penalties [72, 124] (Chapter 2) reach a constant value only in the limit, further extensions would have the be made to accommodate these penalties.

To our knowledge, only Mazumder et al. [124] have developed a principled method to address the issue of tuning parameters for nonconvex penalty. In their work, Mazumder et al. [124] defines the *effective degrees of freedom* based on λ, a and the fit to the data, and reparameterize the penalty to have constant effective degrees of freedom across values of a. So while they do not actually set a or $a\lambda$ explicitly, the reparameterization and constraint has the same general motivation, but without the reference to power calculations.

Given the relevant values, the coefficient value at which the derivative becomes zero is

$$a\lambda = \underset{\beta}{\operatorname{argmin}} \left| \text{power} - p\left(F\left(1, n-2, \frac{n\beta^2 var(\mathbf{x}_j)}{\sigma^2}\right) \geq F_{1-\alpha,1,n-2} \right) \right| \quad (4.48)$$

by combining (4.46) and (4.47). Since λ is known at each step in the regularization path and $a\lambda$ is constant based on (4.48), a is easily calculated.

APPENDIX A

APPENDIX

A.1 Efficient coordinate-wise gradient descent algorithms for high-dimensional penalized generalized linear models

Here we derive the algorithms used by PUMA for efficient estimation in penalized generalized linear models (GLM). In order to make the derivations general so that they are applicable to GLM's with normal or binary responses, we use the following standard notation to describe GLM's as in McCullagh and Nelder [125].

$\mathbf{X}$	matrix of features
$\boldsymbol{\beta}$	vector of regression coefficients
$\mathbf{y}$	vector of observed responses
$\boldsymbol{\eta}^{(s)}$	value of linear predictor at s^{th} iteration
$\ell(\boldsymbol{\eta})$	log-likelihood function of the data based on $\boldsymbol{\eta}$
$\ell(\boldsymbol{\beta})$	reparameterized log-likelihood based on $\boldsymbol{\beta}$
$\boldsymbol{\mu}$	estimated response on the same scale at the observed response, $\mathbf{y}$
$V(\boldsymbol{\mu})$	function of estimated response that is proportional to $var(\mathbf{y})$
$g(.)$	link funtion for a GLM so that $g^{-1}(\boldsymbol{\eta}) = \boldsymbol{\mu}$
θ	cannonical parameter of the GLM
$b(\theta)$	defined so that likelihood is a member of the exponential family
$\mathbf{v}$	$\frac{\partial \ell(\boldsymbol{\eta})}{\partial \boldsymbol{\eta}}$
$\mathbf{W}$	$-\frac{\partial^2 \ell(\boldsymbol{\eta})}{\partial \boldsymbol{\eta} \partial \boldsymbol{\eta}^T}$

A.1.1 Log-likelihood of a generalized linear model

Consider the log-likelihood of a GLM as a function of $\boldsymbol{\eta}$ and obtain the first derivative using

$$\mu_i = b'(\theta_i) = g^{-1}(\boldsymbol{\eta}_i) \tag{A.1}$$

and

$$g(\mu_i) = \boldsymbol{\eta}_i = \mathbf{X}_i\boldsymbol{\beta} \tag{A.2}$$

where vectors are indexed by i [125]. Since the response comes from a distribution in the exponential family, the log-likelihood in cannonical form satisfies

$$\ell(\boldsymbol{\eta}) \quad \propto \quad \sum_{i=1}^{n}[y_i\theta_i - b(\theta_i)] \tag{A.3}$$

The derivative of $\ell(\boldsymbol{\eta})$ follows from the standard properties of a GLM as in McCullagh and Nelder [125]:

$$\begin{aligned}
\frac{\partial \ell}{\partial \eta_i} &= [y_i - b'(\theta_i)]\frac{\partial \theta_i}{\partial \eta_i} && \text{(A.4)}\\
&= [y_i - b'(\theta_i)]\frac{\partial \theta_i}{\partial \mu_i}\frac{\partial \mu_i}{\partial \eta_i} && \text{(A.5)}\\
&= [y_i - b'(\theta_i)]\frac{1}{b''(\theta_i)}\frac{1}{g'(\mu_i)} && \text{(A.6)}\\
&= [y_i - b'(\theta_i)]\frac{1}{V(\mu_i)g'(\mu_i)} && \text{(A.7)}\\
&= \frac{(y_i - \mu_i)}{V(\mu_i)g'(\mu_i)} && \text{(A.8)}
\end{aligned}$$

Using the fact that $V(\mu_i)g'(\mu_i) = 1$ if the cannonical link is used and expressing the derivative in matrix notation:

$$\begin{aligned}
\mathbf{v}_{n\times 1} &= \frac{\partial \ell}{\partial \boldsymbol{\eta}} && \text{(A.9)}\\
&= \mathbf{y} - \boldsymbol{\mu} && \text{(A.10)}
\end{aligned}$$

The second derivative is:

$$\frac{\partial^2 \ell}{\partial \eta_i \partial \eta_{i'}} = \frac{\partial \mu_i}{\partial \eta_{i\prime}} \tag{A.11}$$

$$= \frac{1}{g'(\mu_{i\prime})} \tag{A.12}$$

Let $\mathbf{W}$ denote the negative Hessian and notice that the off-diagonal elements are zero:

$$\mathbf{W}_{n\times n} = -\frac{\partial \ell}{\partial \boldsymbol{\eta} \partial \boldsymbol{\eta}^T} \tag{A.13}$$

$$= \text{diag}\left[\frac{1}{g'(\boldsymbol{\mu})}\right] \tag{A.14}$$

If the cannonical link is used, the formula for $\mathbf{W}$ is:

response	$g(\mu_i)$	$g'(\mu_i)$	$\mathbf{W}$
normal	1	1	diag(1)
binomial	$\log \frac{\mu_i}{1-\mu_i}$	$\frac{1}{\mu_i(1-\mu_i)}$	diag $[\boldsymbol{\mu}(1-\boldsymbol{\mu})]$

A.1.2 Quadratic approximation to the log-likelihood

The standard iteratively reweighted least squares (IRLS) algorithm follows from a quadratic approximation of the log-likelihood. We derive this approximation here and derive the iterative algorithm in the next section. Based on the second order Taylor expansion of the log-likelihood,

$$\ell(\boldsymbol{\eta}) \approx \ell(\boldsymbol{\eta}^{(s)}) + (\boldsymbol{\eta} - \boldsymbol{\eta}^{(s)})^T \ell'(\boldsymbol{\eta}^{(s)}) + \frac{1}{2}(\boldsymbol{\eta} - \boldsymbol{\eta}^{(s)}) \ell''(\boldsymbol{\eta}^{(s)})(\boldsymbol{\eta} - \boldsymbol{\eta}^{(s)}) \tag{A.15}$$

$$= (\boldsymbol{\eta} - \boldsymbol{\eta}^{(s)})^T \mathbf{v} - \frac{1}{2}(\boldsymbol{\eta} - \boldsymbol{\eta}^{(s)})^T W (\boldsymbol{\eta} - \boldsymbol{\eta}^{(s)}) \tag{A.16}$$

$$= -\frac{1}{2}(\boldsymbol{\eta} - \boldsymbol{\eta}^{(s)} - \mathbf{v}\mathbf{W}^{-1})^T \mathbf{W} (\boldsymbol{\eta} - \boldsymbol{\eta}^{(s)} - \mathbf{v}\mathbf{W}^{-1}) \tag{A.17}$$

$$= -\frac{1}{2}(\tilde{\mathbf{y}} - \boldsymbol{\eta})^T \mathbf{W} (\tilde{\mathbf{y}} - \boldsymbol{\eta}) \tag{A.18}$$

where the working response, $\tilde{\mathbf{y}} = \boldsymbol{\eta}^{(s)} + \mathbf{v}\mathbf{W}^{-1}$, is a function of the linear predictor at the s^{th} iteration. Reparameterizing as a function of $\boldsymbol{\beta}$,

$$\ell(\boldsymbol{\beta}) \quad = \quad -\frac{1}{2}(\tilde{\mathbf{y}} - \mathbf{X}\boldsymbol{\beta})^T\mathbf{W}(\tilde{\mathbf{y}} - \mathbf{X}\boldsymbol{\beta}) \tag{A.19}$$

This is equivalent to the IRLS Fisher scoring system where the Hessian is replaced by its expectation [125]. In standard Fisher scoring notation, the expected Hessian is:

$$E\left[\frac{\partial^2 \ell}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}}\right] \quad = \quad -\mathbf{X}^T\mathbf{W}\mathbf{X} \tag{A.20}$$

as is the case in equation (A.19). Expressed in stardard notation, the working responses are

$$\tilde{\mathbf{y}} = g(\boldsymbol{\mu}) + g'(\boldsymbol{\mu})(\mathbf{y} - \boldsymbol{\mu}) \tag{A.21}$$

where it is trivial to show this is equivalent to $\tilde{\mathbf{y}} = \boldsymbol{\eta}^{(s)} + \mathbf{v}\mathbf{W}^{-1}$.

Details of derivation

Equation numbers and an explanation of their derivation:

(A.15) A second order approximation to $\ell(\boldsymbol{\eta})$ as in Breheny and Huang [16]. This is equivalent to the Fisher scoring algorithm of IRLS.

(A.16) Plug in $\mathbf{v}$ for the first derivative and $-\mathbf{W}$ for the Hessian and drop terms that don't depend on $\boldsymbol{\eta}$ (i.e. drop $\ell(\boldsymbol{\eta}^{(s)})$)

(A.17) Complete the square as in Breheny and Huang [16]: In general, $ax^2 + bx + c = a(x-h)^2 + k$ where $h = -\frac{b}{2a}$ and $k = c - \frac{b^2}{4a}$ In this case, $a = -\frac{1}{2}\mathbf{W}$, $b = \mathbf{v}$, $c = 0$ and $x = (\mathbf{X}\boldsymbol{\beta} - \boldsymbol{\eta}^{(s)})$ so $h = \mathbf{v}\mathbf{W}^{-1}$ and k can be dropped since it does not depend on $\boldsymbol{\beta}$.

(A.18) Plug in $\tilde{\mathbf{y}} = \boldsymbol{\eta}^{(s)} + \mathbf{v}\mathbf{W}^{-1}$

A.1.3 Estimation in unpenalized generalized linear models

While standard IRLS methods update all regression parameters in a single step, here we derive a coordinate-wise gradient descent algorithm for estimating $\boldsymbol{\beta}$ in an unpenalized GLM and extend it to the penalized case in the next section. Following Breheny and Huang [17], Friedman et al. [47, 48], consider the quadratic approximation to the log-likelihood and construct a coordinate-wise Newton-Raphson update of β_j using the first and second derivatives.

$$\begin{aligned}
\ell(\boldsymbol{\beta}) &= -\frac{1}{2}(\tilde{\mathbf{y}} - \mathbf{X}\boldsymbol{\beta})^T\mathbf{W}(\tilde{\mathbf{y}} - \mathbf{X}\boldsymbol{\beta}) && \text{(A.22)}\\
&= -\frac{1}{2}(\tilde{\mathbf{y}} - \sum_{k\neq j} x_k\beta_k - x_j\beta_j)^T\mathbf{W}(\tilde{\mathbf{y}} - \sum_{k\neq j}\mathbf{x}_k\beta_k - \mathbf{x}_j\beta_j) && \text{(A.23)}\\
\frac{\partial \ell}{\partial \beta_j} &= (\tilde{\mathbf{y}} - \sum_{k\neq j}\mathbf{x}_k\beta_k - \mathbf{x}_j\beta_j)^T\mathbf{W}\mathbf{x}_j && \text{(A.24)}\\
&= (\tilde{\mathbf{y}} - \sum_{k\neq j}\mathbf{x}_k\beta_k)^T\mathbf{W}\mathbf{x}_j - \beta_j\mathbf{x}_j^T\mathbf{W}\mathbf{x}_j && \text{(A.25)}\\
&= (\mathbf{x}_j\beta_j^{(s)} + \mathbf{v})^T\mathbf{W}\mathbf{x}_j - \beta_j\mathbf{x}_j^T\mathbf{W}\mathbf{x}_j && \text{(A.26)}\\
&= \beta_j^{(s)}\mathbf{x}_j^T\mathbf{W}\mathbf{x}_j + \mathbf{v}^T\mathbf{W}\mathbf{x}_j - \beta_j\mathbf{x}_j^T\mathbf{W}\mathbf{x}_j && \text{(A.27)}\\
\frac{\partial^2 \ell}{\partial\beta_j\partial\beta_j} &= -\mathbf{x}_j^T\mathbf{W}\mathbf{x}_j && \text{(A.28)}
\end{aligned}$$

The coordinate-wise update follows from the standard Newton-Raphson update formula.

$$\begin{aligned}
\beta_j^{(s+1)} &= \beta_j^{(s)} - \frac{\ell'(\beta_j^{(s)})}{\ell''(\beta_j^{(s)})} && \text{(A.29)}\\
&= \beta_j^{(s)} + \frac{\beta_j^{(s)}\mathbf{x}_j^T\mathbf{W}\mathbf{x}_j + \mathbf{v}^T\mathbf{W}\mathbf{x}_j - \beta_j^{(s)}\mathbf{x}_j^T\mathbf{W}\mathbf{x}_j}{\mathbf{x}_j^T\mathbf{W}\mathbf{x}_j} && \text{(A.30)}\\
&= \beta_j^{(s)} + \frac{\mathbf{v}^T\mathbf{W}\mathbf{x}_j}{\mathbf{x}_j^T\mathbf{W}\mathbf{x}_j} && \text{(A.31)}
\end{aligned}$$

This gives the coordinate-wise update:

$$\boxed{\beta_j^{(s+1)} = \frac{\beta_j^{(s)}\mathbf{x}_j^T\mathbf{W}\mathbf{x}_j + \mathbf{v}^T\mathbf{W}\mathbf{x}_j}{\mathbf{x}_j^T\mathbf{W}\mathbf{x}_j}} \tag{A.32}$$

A.1.4 Estimation in penalized generalized linear models

Following the derivation from the previous section, now consider a penalized GLM where a penalty term is added to the standard log-likelihood of a GLM. Specifically, consider a GLM log-likelihood with an arbitrary penalty $p(.)$ on the magnitute of β_j:

$$\ell(\boldsymbol{\beta}) = -\frac{1}{2}(\tilde{\mathbf{y}} - \mathbf{X}\beta)^T\mathbf{W}(\tilde{\mathbf{y}} - \mathbf{X}\beta) - \sum_j p(\beta_j) \tag{A.33}$$

The coordinate-wise Newton-Raphson update is:

$$\beta_j^{(s+1)} = \beta_j^{(s)} - \frac{\ell'(\beta_j^{(s)}) - p'(\beta_j^{(s)})}{\ell''(\beta_j^{(s)}) - p''(\beta_j^{(s)})} \tag{A.34}$$

$$= \beta_j^{(s)} + \frac{\mathbf{v}^T\mathbf{W}\mathbf{x}_j - p'(\beta_j^{(s)})}{\mathbf{x}_j^T\mathbf{W}\mathbf{x}_j + p''(\beta_j^{(s)})} \tag{A.35}$$

While this update considers $p(.)$ in general, specific penalty functions are considered below.

Least Absolute Shrinkage and Selection Operator (LASSO)

Proposed by Tibshirani [187]. Consider the penalty function and its derivatives.

$$p_{\text{LASSO}}(\beta_j) = \lambda|\beta_j| \tag{A.36}$$

$$p'_{\text{LASSO}}(\beta_j) = \lambda sign(\beta_j) \tag{A.37}$$

$$p''_{\text{LASSO}}(\beta_j) = 0 \tag{A.38}$$

The coordinate-wise Newton-Raphson update is

$$\begin{aligned}
\beta_j^{(s+1)} &= \beta_j^{(s)} + \frac{\mathbf{v}^T\mathbf{W}\mathbf{x}_j - \lambda sign(\beta_j^{(s)})}{\mathbf{x}_j^T\mathbf{W}\mathbf{x}_j} && \text{(A.39)} \\
&= \frac{\beta_j^{(s)}\mathbf{x}_j^T\mathbf{W}\mathbf{x}_j + \mathbf{v}^T\mathbf{W}\mathbf{x}_j - \lambda sign(\beta_j^{(s)})}{\mathbf{x}_j^T\mathbf{W}\mathbf{x}_j} && \text{(A.40)} \\
&= \frac{d - \lambda sign(d)}{\mathbf{x}_j^T\mathbf{W}\mathbf{x}_j} && \text{(A.41)}
\end{aligned}$$

where $d = \beta_j^{(s)}\mathbf{x}_j^T\mathbf{W}\mathbf{x}_j + \mathbf{v}^T\mathbf{W}\mathbf{x}_j$.

Thus $|d| > \lambda$ yields $(|d| - \lambda)sign(d)$ but $|d| \leq \lambda$ yields an update of 0 so that $\beta_j^{(s+1)}$ is set to zero if the derivative of the lasso penalty exceeds the unpenalized update. This can be stated using the soft-thresholding function of Donoho and Johnstone [35] (see Tibshirani [187] for orginal use in lasso regression):

$$\boxed{\beta_j = \frac{S(d, \lambda)}{\mathbf{x}_j^T\mathbf{W}\mathbf{x}_j}} \tag{A.42}$$

where

$$S(d, \lambda) = \begin{cases} d - \lambda & d > 0 \text{ and } \lambda < |d| \\ d + \lambda & d < 0 \text{ and } \lambda < |d| \\ 0 & \lambda \geq |d| \end{cases} \tag{A.43}$$

Note the similar application of the soft-thresholding function by Friedman et al. [47, 48], Schifano et al. [166], Wu et al. [209], Wu and Lange [210]. Also note that $\ell(\boldsymbol{\beta})$ is not differentiable when $\beta_j = 0$, but a directional derivative still exists and has magnitude λ [210]. Therefore the updates above are valid even if $\beta_j = 0$.

Minimax concave penalty (MCP)

Proposed by Zhang [221]. Consider the penalty function and its derivatives.

$$p_{\text{MCP}}(\beta_j) = \begin{cases} \lambda|\beta_j| - \frac{\beta_j^2}{2a} & |\beta_j| \leq a\lambda \\ \frac{1}{2}a\lambda^2 & |\beta_j| > a\lambda \end{cases} \tag{A.44}$$

$$p'_{\text{MCP}}(\beta_j) = \begin{cases} \lambda sign(\beta_j^{(s)}) - \frac{\beta_j}{a} & |\beta_j| \leq a\lambda \\ 0 & |\beta_j| > a\lambda \end{cases} \tag{A.45}$$

$$p''_{\text{MCP}}(\beta_j) = \begin{cases} -\frac{1}{a} & |\beta_j| \leq a\lambda \\ 0 & |\beta_j| > a\lambda \end{cases} \tag{A.46}$$

The coordinate-wise Newton-Raphson update is

$$\beta_j^{(s+1)} = \beta_j^{(s)} + \frac{\mathbf{v}^T\mathbf{W}\mathbf{x}_j - \lambda sign(\beta_j^{(s)}) + \frac{\beta_j^{(s)}}{a}}{\mathbf{x}_j^T W\mathbf{x}_j - \frac{1}{a}} \tag{A.47}$$

$$= \frac{\beta_j^{(s)}(\mathbf{x}_j^T\mathbf{W}\mathbf{x}_j - \frac{1}{a}) + \mathbf{v}^T\mathbf{W}\mathbf{x}_j - \lambda sign(\beta_j^{(s)}) + \frac{\beta_j^{(s)}}{a}}{\mathbf{x}_j^T\mathbf{W}\mathbf{x}_j - \frac{1}{a}} \tag{A.48}$$

$$= \frac{S(d,\lambda)}{\mathbf{x}_j^T\mathbf{W}\mathbf{x}_j - \frac{1}{a}} \tag{A.49}$$

$$\boxed{\beta_j^{(s+1)} = \begin{cases} \frac{S(d,\lambda)}{\mathbf{x}_j^T\mathbf{W}\mathbf{x}_j - \frac{1}{a}} & d \leq a\lambda\mathbf{x}_j^T\mathbf{W}\mathbf{x}_j \\ \frac{d}{\mathbf{x}_j^T\mathbf{W}\mathbf{x}_j} & d > a\lambda\mathbf{x}_j^T\mathbf{W}\mathbf{x}_j \end{cases}} \tag{A.50}$$

Notice the bounds change in (A.50) to adjust for the scale of the second derivative according to Breheny and Huang [16, 17].

Negative exponential gamma (NEG)

Proposed by Griffin and Brown [54, 55], but see Hoggart et al. [72] for more details. Consider the penalty function and its derivatives.

$$p_{\text{NEG}}(\beta) = -\frac{\beta^2}{4\gamma^2} - \log D_{-2\lambda-1}(\frac{|\beta|}{\gamma}) \tag{A.51}$$

$$p'_{\text{NEG}}(\beta) = sign(\beta)\frac{2\lambda+1}{\gamma}\frac{D_{-2\lambda-2}(\frac{|\beta|}{\gamma})}{D_{-2\lambda-1}(\frac{|\beta|}{\gamma})} \tag{A.52}$$

$$p''_{\text{NEG}}(\beta_j) = -\frac{4}{\gamma}\left[(\lambda+1)(\lambda+\frac{1}{2})\frac{D_{-2\lambda-3}(\frac{|\beta|}{\gamma})}{D_{-2\lambda-1}(\frac{|\beta|}{\gamma})} - \left[(\lambda+\frac{1}{2})\frac{D_{-2\lambda-2}(\frac{|\beta|}{\gamma})}{D_{-2\lambda-1}(\frac{|\beta|}{\gamma})}\right]^2\right] \tag{A.53}$$

$D_{\cdot}(.)$ denotes the parabolic cylinder function [139]. We use a Fortran implentation of this function available in the SciPy library: `http://projects.scipy.org/scipy/export/6949/trunk/scipy/special/specfun/specfun.f`

The coordinate-wise Newton-Raphson update is

$$\boxed{\beta_j = \beta_j^{(s)} + \frac{\mathbf{v}^T\mathbf{W}\mathbf{x}_j - p'_{\text{NEG}}(\beta_j^{(s)})}{\mathbf{x}_j^T\mathbf{W}\mathbf{x}_j - p''_{\text{NEG}}(\beta_j^{(s)})}} \tag{A.54}$$

and cannot be simplified due to the complicated form for the penalty.

LOG penalty

Proposed by Mazumder et al. [124]. Consider the penalty function and its derivatives.

$$p_{\text{LOG}}(\beta_j) = \lambda\frac{\log(1+|\beta_j|/\epsilon)}{\log(1+1/\epsilon)} \tag{A.55}$$

$$p'_{\text{LOG}}(\beta_j) = \frac{\lambda}{\log(1+1/\epsilon)}\frac{sign(\beta_j)/\epsilon}{1+|\beta_j|/\epsilon} \tag{A.56}$$

$$p''_{\text{LOG}}(\beta_j) = \frac{-\lambda}{\log(1+1/\epsilon)}\frac{[sign(\beta_j)/\epsilon]^2}{[1+|\beta_j|/\epsilon]^2} \tag{A.57}$$

The coordinate-wise Newton-Raphson update is

$$\boxed{\beta_j = \beta_j^{(s)} + \frac{\mathbf{v}^T \mathbf{W} \mathbf{x}_j - p'_{\text{LOG}}(\beta_j^{(s)})}{\mathbf{x}_j^T \mathbf{W} \mathbf{x}_j - p''_{\text{LOG}}(\beta_j^{(s)})}} \tag{A.58}$$

A.1.5 Updating quantities during iteration of algorithm

When β_j is updated, $\mathbf{v}$ and $\mathbf{W}$ must be updated accordingly. In general, initialize $\boldsymbol{\eta}$ as:

$$\boldsymbol{\eta}^{(0)} = \mathbf{X}\boldsymbol{\beta}^{(0)}. \tag{A.59}$$

At each iteration where $\beta_j^{(s+1)} \neq \beta_j^{(s)}$, update $\boldsymbol{\eta}$, $\mathbf{v}$ and $\mathbf{W}$ according to

$$\boldsymbol{\eta}^{(s+1)} = \boldsymbol{\eta}^{(s)} + \mathbf{x}_j \left(\beta_j^{(s+1)} - \beta_j^{(s)} \right) \tag{A.60}$$

$$\boldsymbol{\mu} = g^{-1}(\eta) \tag{A.61}$$

$$\mathbf{v} = \mathbf{y} - \boldsymbol{\mu} \tag{A.62}$$

$$\mathbf{W} = 1/g'(\boldsymbol{\mu}) \tag{A.63}$$

A.1.6 Convergence and a minorize-maximization algorithm

It is widely known that Newton-Raphson algorithms are not guaranteed to converge unless the objective function is quadratic. Since we are using a quadratic approximation of the objective function, the Fisher scoring system is not guaranteed to converge. The Fisher scoring method of IRLS tends to converge in practice in overdetermined systems, yet the highly underdetermined systems addressed here may be problematic. In fact, Hoggart et al. [72] and Genkin et al. [51] replace the Hessian with an upper bound so that the system always converges. Intuitively, this upper bound prevents taking steps that are too large

and actually decrease the value of the objective function.

These examples are generalized by Hunter et al. [75], who describe a very general method to construct a surrogate function that minorizes the objective function. They demonstrate that iteratively maximizing the surrogate and constructing a new surrogate can guarantee convergence on non-quadratic surfaces without expensive backtracking checks that evaluate the log-likelihood at each update. One simple method to construct a surrogate function which minorizes the objective function is to replace the Hessian with an upper bound. This effectively decreases the step size so that the value of the objective function always increases and a backtracking step is not needed.

Convergence can therefore be guaranteed if the matrix $\mathbf{W}$ used in the quadratic approximation is replaced by an upper bound. Consider the upper bound on the Hessian for the following links:

- linear: the system is exactly quadratic so that $\mathbf{W} = \text{diag}(1)$ and no bound is needed.
- logistic: $\mathbf{W} = \text{diag}(\boldsymbol{\mu}(1-\boldsymbol{\mu}))$ so that $\mathbf{W}_{upper} = \frac{1}{4}\text{diag}(1)$
  since $\mu_i \in [0,1]$ and $\max(\mu_i(1-\mu_i)) = .25$
  (see example in Hunter et al. [75])

For logistic regression, replacing $\mathbf{W}$ with $\mathbf{W}_{upper}$ to guarantees convergence and increases the speed of each update since it avoids calculating $\boldsymbol{\mu}$ and $\boldsymbol{\mu}(1-\boldsymbol{\mu})$.

A.1.7 Implementation

Our implementation of these penalized likelihood methods is especially efficient since we 1) store the dataset to mimimize access time for each feature, 2) use highly optimized Intel® Math Kernel Library® for linear algebra operations and 3) evaluate multiple modes of the nonconvex posterior surface in parallel using OpenMP®. This very fast implementation allowed efficient exploration of the two-dimensional space of tuning parameters for MCP, LOG and NEG penalties using the relaxation method of Mazumder et al. [124].

A.2 Running HyperLasso

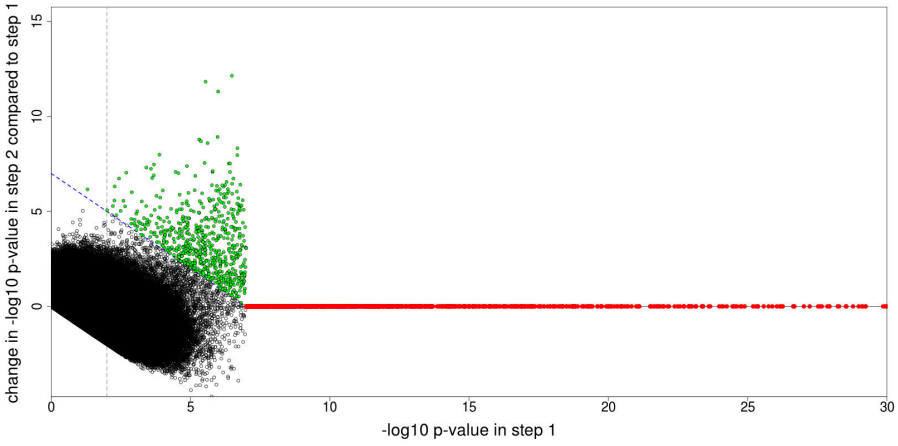
We ran hyperLasso following the instructions on the program website (`http://www.ebi.ac.uk/projects/BARGEN/`) that suggested the following flags:

```
-shape 0.1 -lambda 50 -std
```

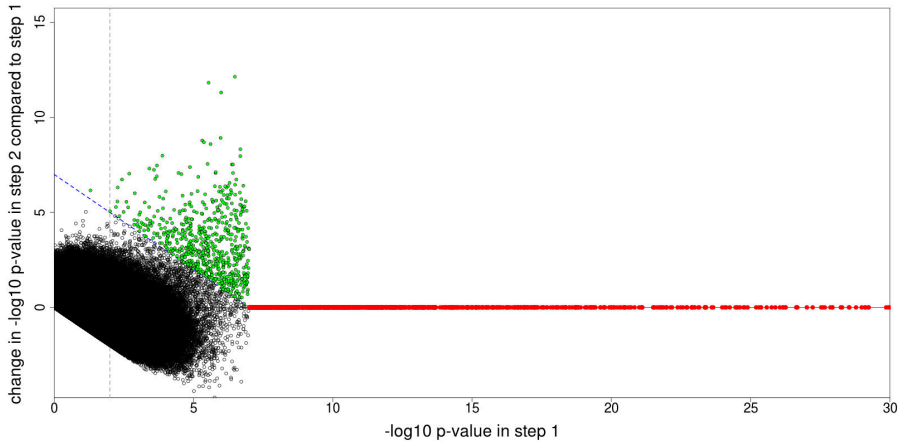
Setting the lambda values based on asymptotic arguments described by Hoggart et al. [72] either produced very few markers with non-zero coefficients or caused the program to crash.

Figure A.1: Assessing p-value cutoff in two-step forward regression. Plots show $-\log_{10}$ p-values from a single marker analysis (x-axis) compared to the change in $-\log_{10}$ p-values from a conditional regression analysis where markers passing the Bonferroni cutoff are included as covariates (y-axis). Markers passing the Bonferroni cutoff in the first step (red points) are necessary omitted from being tested in the second step, and are considered to have no change in p-value. Markers with a large enough increase in $-\log_{10}$ p-value in the second step to cross the second Bonferroni cutoff (blue dashed line) are indicated by green points. The p-value cutoff of 0.01 (i.e. a $-\log_{10}$ p-value of 2) is indicated by the grey dashed line. Results are shown for 10 replicate simulations each of (a) 1000, (b) 2000, and (c) 5000 samples with 500K markers, heritability of 30, 40, 50 or 60% and 30, 40, 50, 70 or 100 simulated markers with true nonzero coefficients. This corresponds to 200 simulations and 1×10^8 p-values for each sample size. The results indicate that in a forward regression, which approximates penalized multiple regression [37, 63], markers with small $-\log_{10}$ p-values in the first step have a very low probability of being significant in the second step. Therefore, using a p-value cutoff of 0.01 from a marginal regression retains almost all relevant variables under biologically motivated simulation conditions.

(a) 1000 samples



(b) 2000 samples



(c) 5000 samples

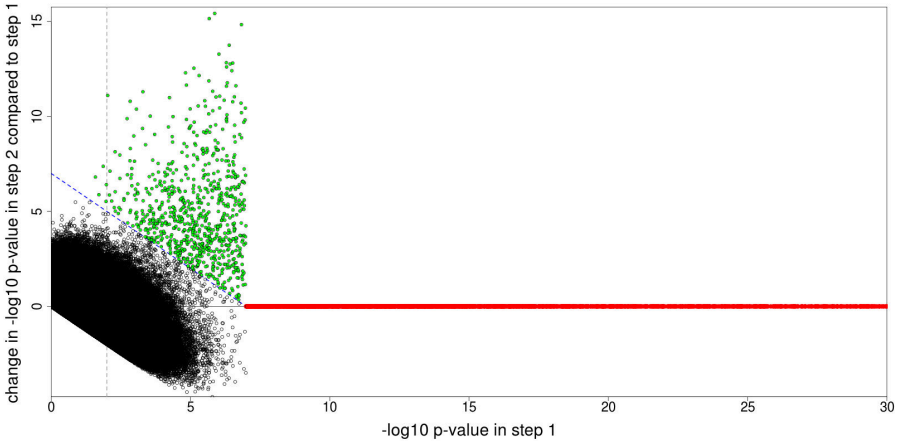
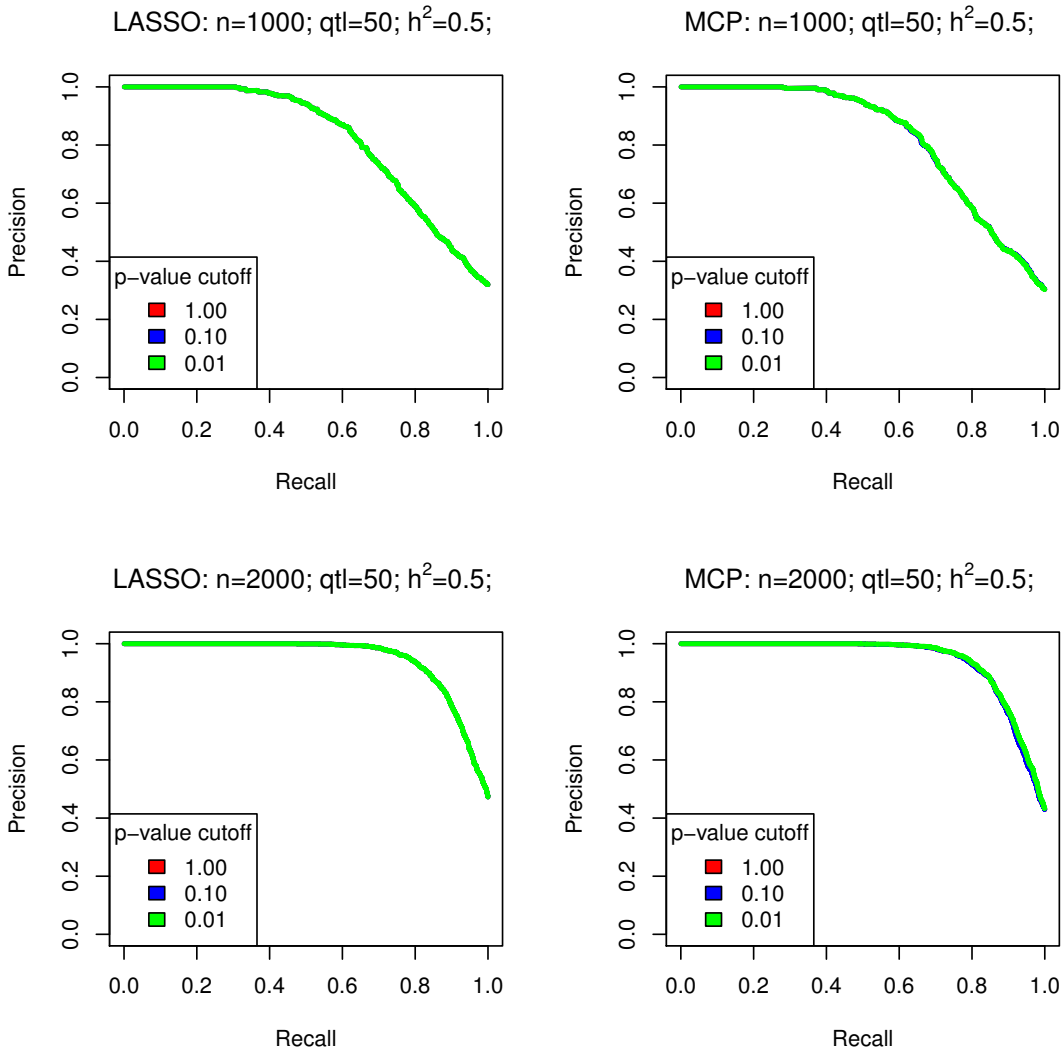


Figure A.2: Effect of pre-screening on performance of PUMA. (a) 100 replicate simulations with 500K markers, 50 causal markers, heritability of 50% and 1000 or 2000 individuals using Lasso and MCP methods show that using a pre-screening p-value cutoff of 1, 0.10 and 0.01 has no noticeable effect of performance of PUMA. Note that the performance was so similar for all cutoffs that the curves are overlapping. (b) Running times for simulations in (a) show that pre-screening substantually reduces computational time. We note that simulations with 5000 individuals were not possible due to the very high memory requirements of running PUMA without prescreening.

(a) Precision-Recall curves



(b) Run times for each prescreening cutoff

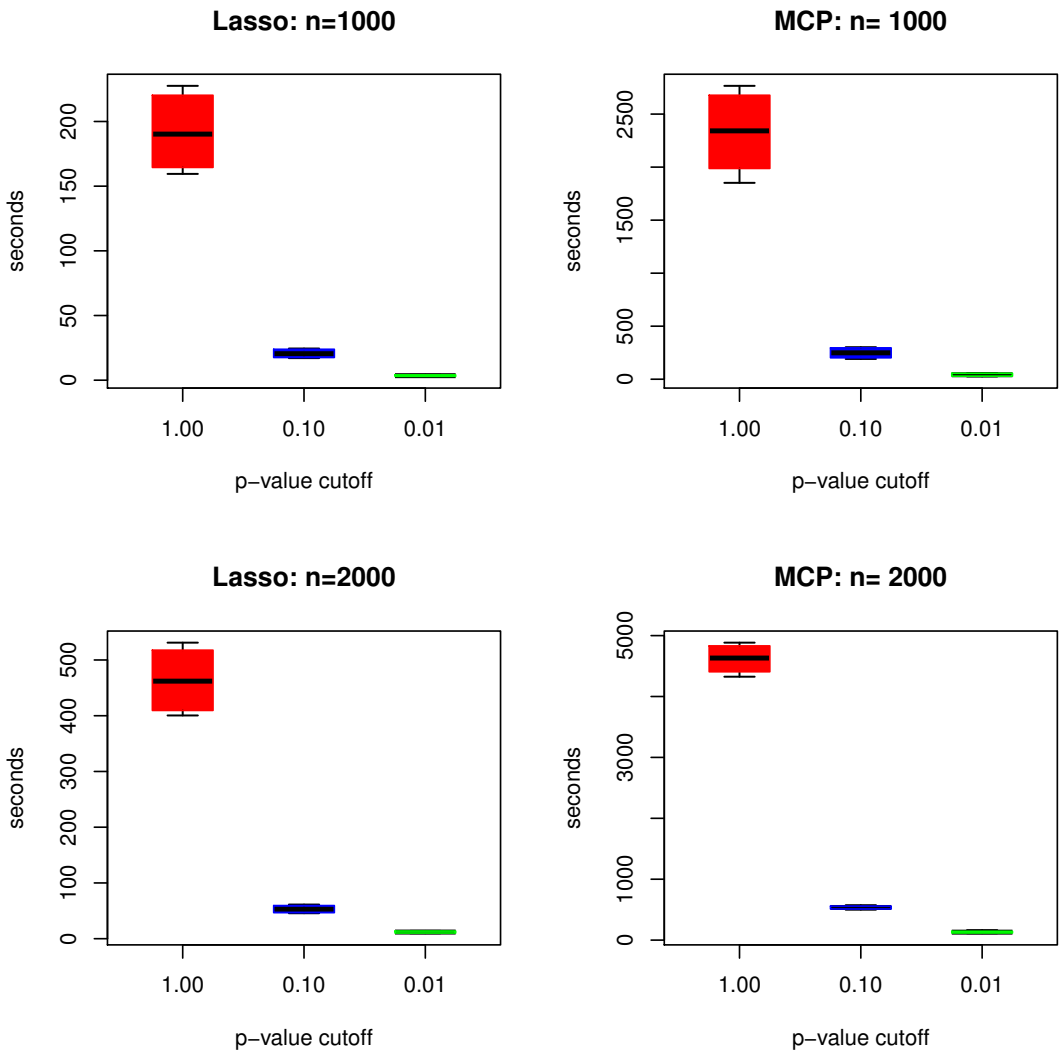


Figure A.3: Simulation results showing power for our PMR methods, current PMR methods, an approximate Bayesian method, single marker analysis and conditional regression methods at an FDR of 5% as a function of sample size as in Figure 2.3c in the main text. Results are shown for a range of total heritabilities and number of susceptibility loci.

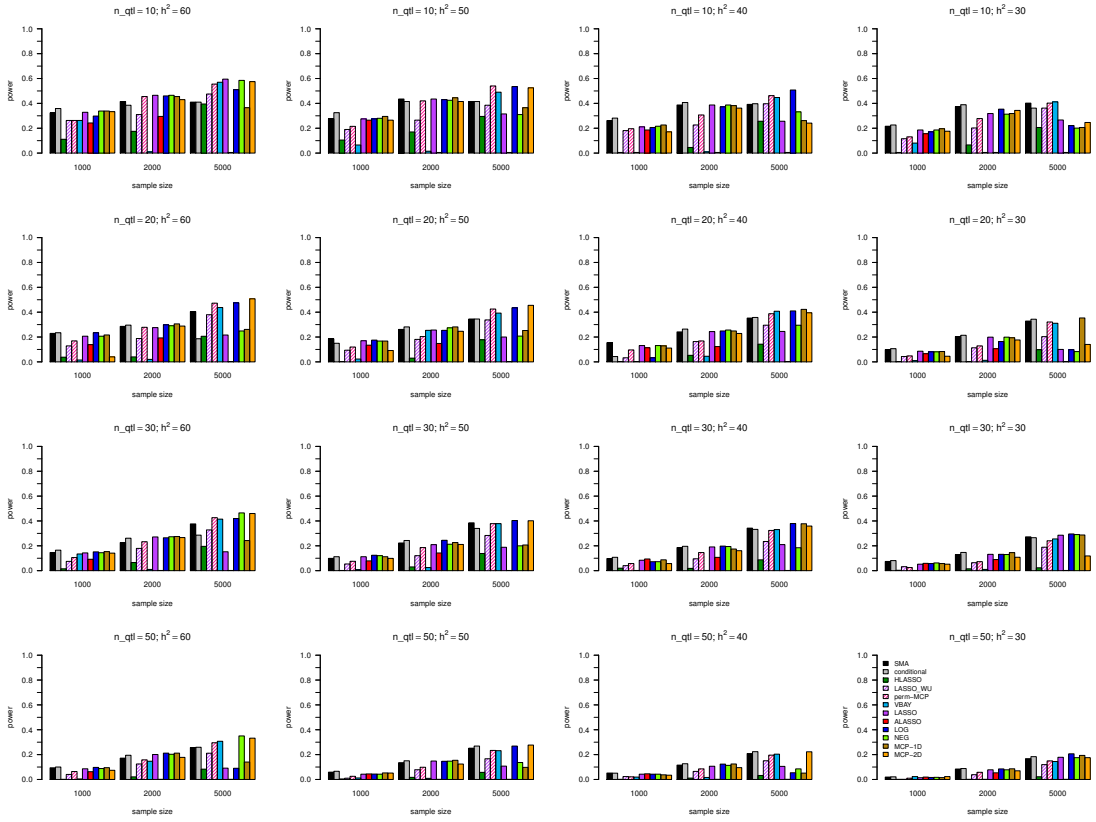


Figure A.4: Simulation results showing power for our PMR methods, current PMR methods, an approximate Bayesian method, single marker analysis and conditional regression methods at an FDR of 5% as a function of the number of susceptibility loci. Results are shown for a range of sample sizes and total heritabilities.

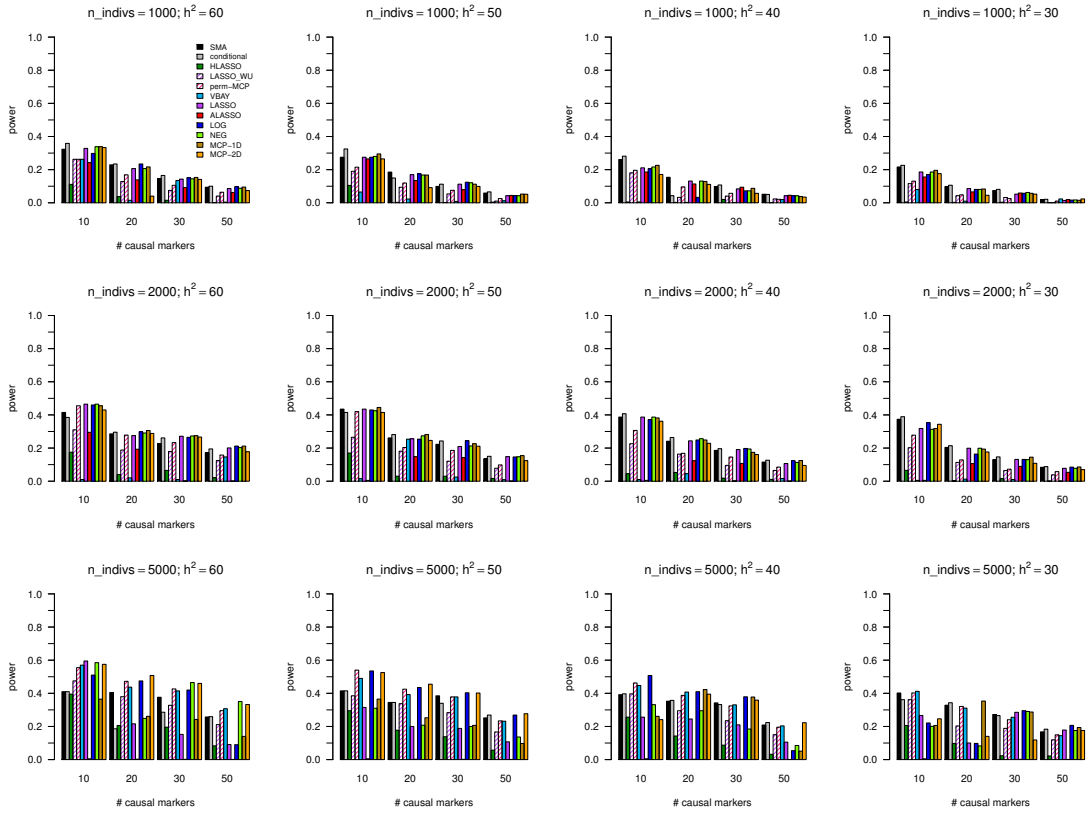


Figure A.5: Simulation results showing power a sample size of 1000 for our PMR methods, current PMR methods, an approximate Bayesian method, single marker analysis and conditional regression methods at an FDR of 5% as a function of the marginal heritability of each causal marker as in Figure 2.3a in the main text. Results are shown for a range of total heritabilities and number of causal markers.

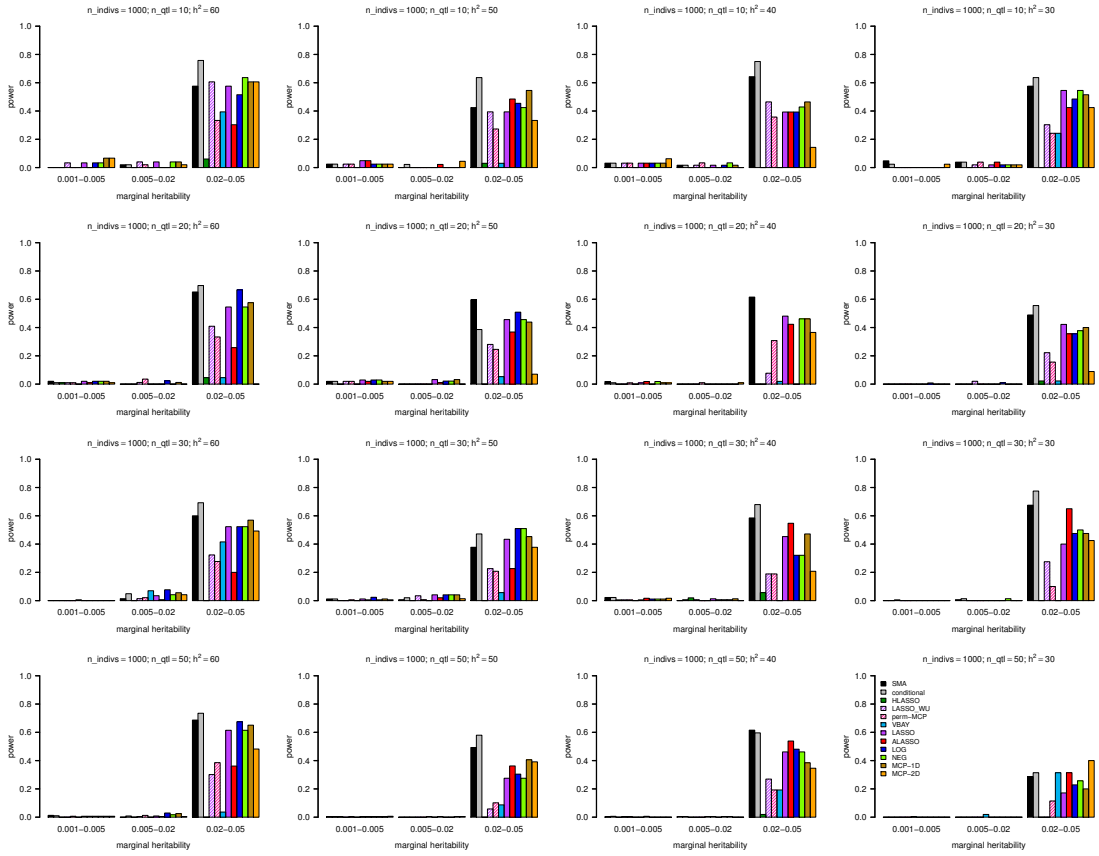


Figure A.6: Simulation results showing power a sample size of 2000 for our PMR methods, current PMR methods, an approximate Bayesian method, single marker analysis and conditional regression methods at an FDR of 5% as a function of the marginal heritability of each causal marker as in Figure 2.3a in the main text. Results are shown for a range of total heritabilities and number of causal markers..

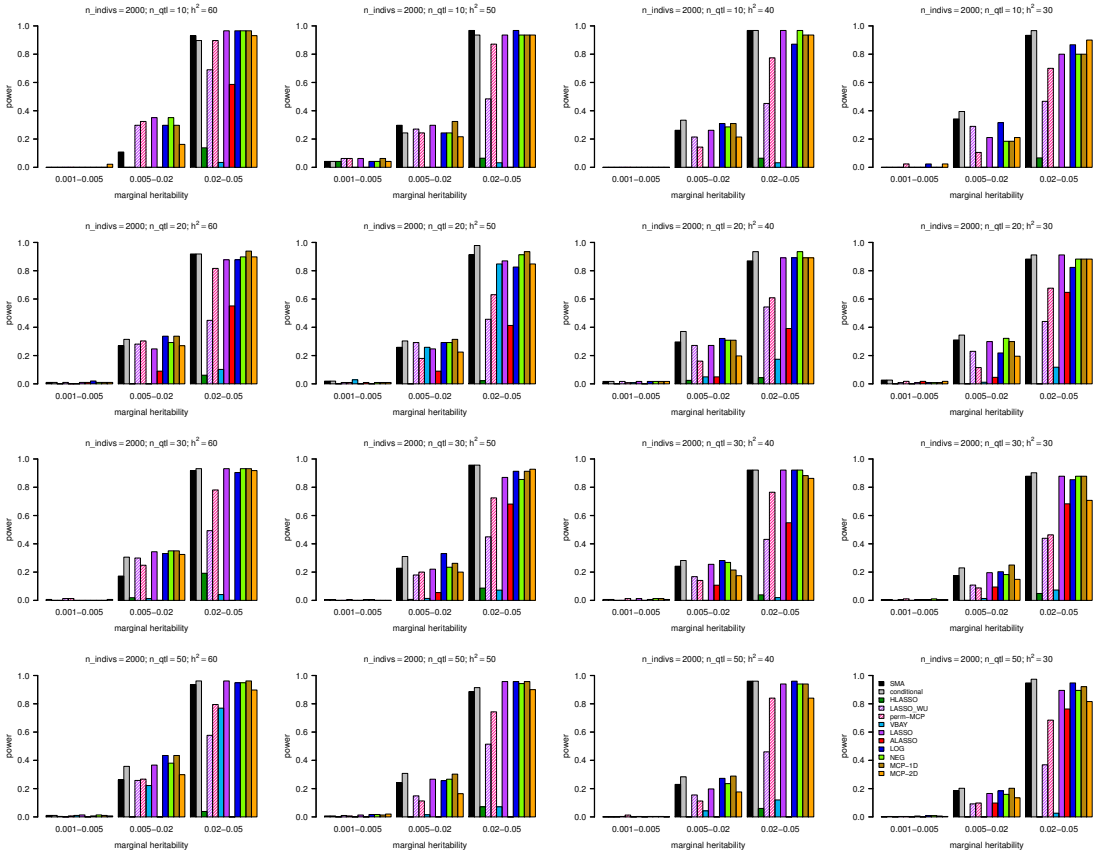


Figure A.7: Simulation results showing power a sample size of 5000 for our PMR methods, current PMR methods, an approximate Bayesian method, single marker analysis and conditional regression methods at an FDR of 5% as a function of the marginal heritability of each causal marker as in Figure 2.3a in the main text. Results are shown for a range of total heritabilities and number of causal markers.

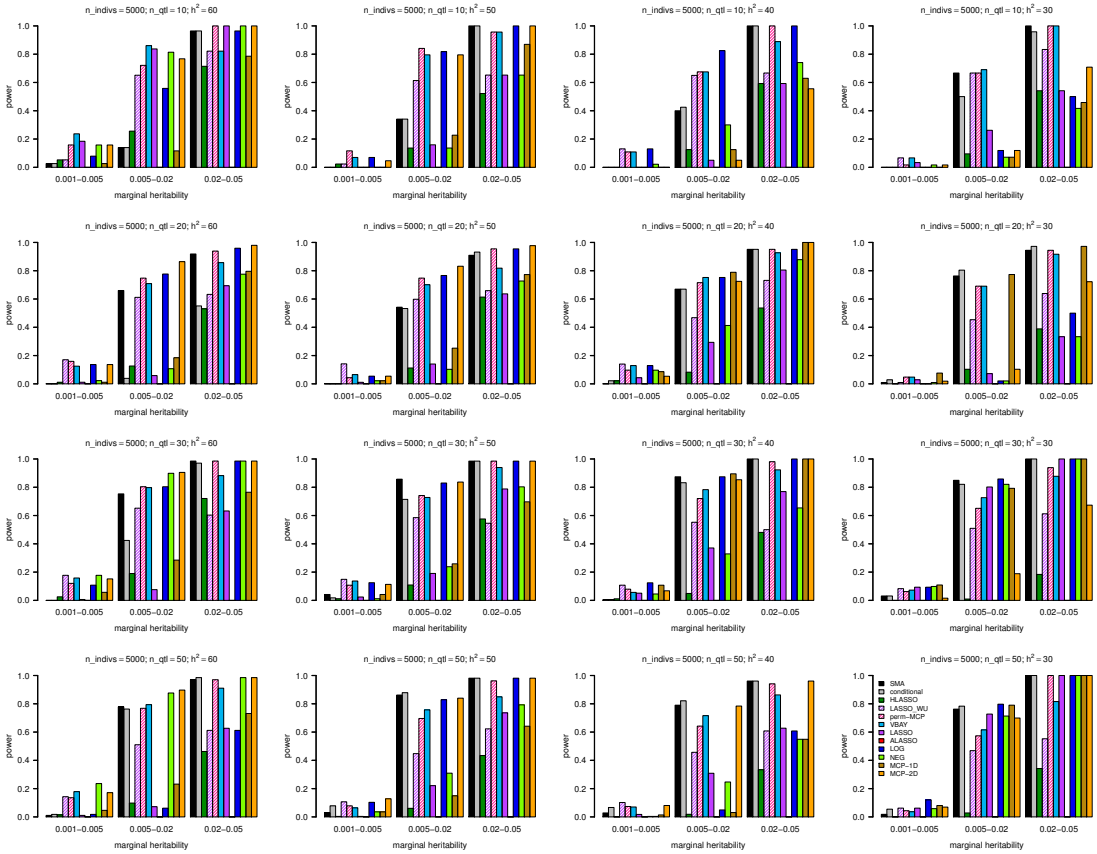


Figure A.8: Simulation results showing precision-recall curves for a sample size of 1000. Results are shown for a range of total heritabilities and number of causal markers. Solid colors for pML methods indicate results using our method for assessing significance in the presence of correlated markers, while dashes indicate the significance method of Wu et al. [209] and perm-MCP [9]. This figures is analogous to Figure 2.3b in the main text.

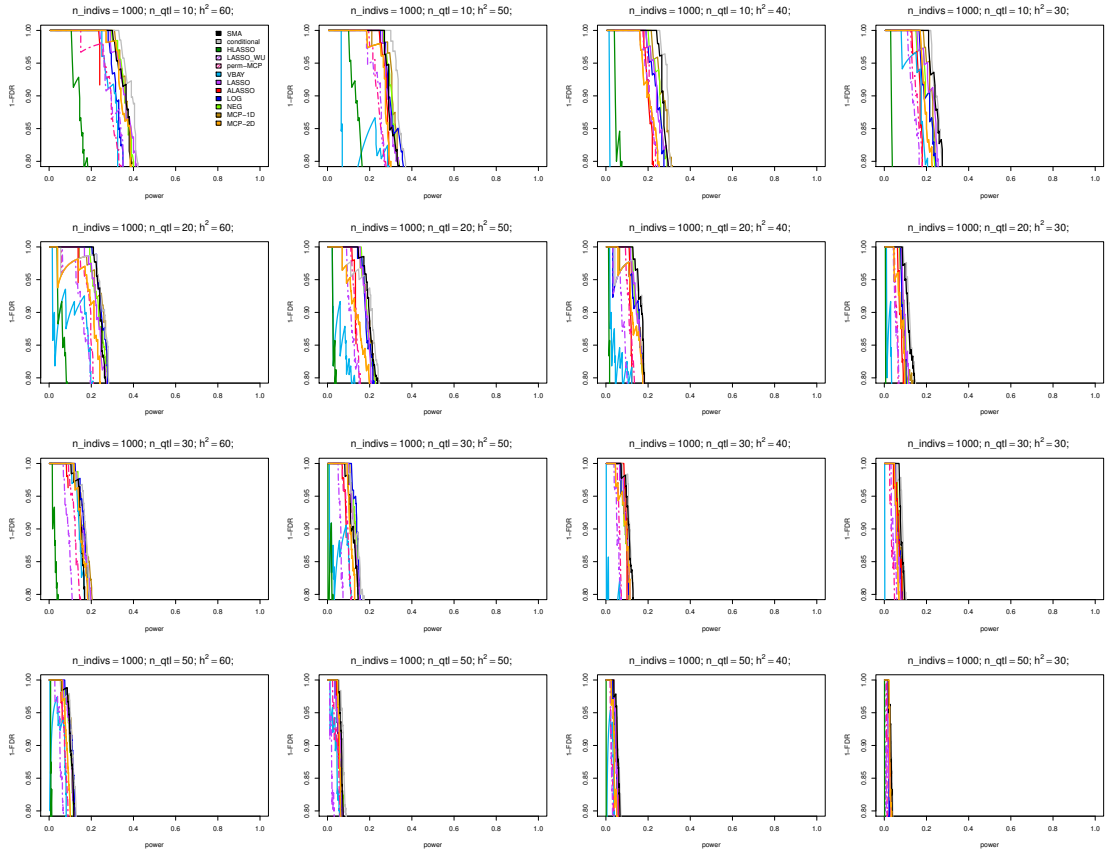


Figure A.9: Simulation results showing precision-recall curves for a sample size of 2000. Results are shown for a range of total heritabilities and number of causal markers. Solid colors for pML methods indicate results using our method for assessing significance in the presence of correlated markers, while dashes indicate the significance method of Wu et al. [209] and perm-MCP [9]. This figures is analogous to Figure 2.3b in the main text.

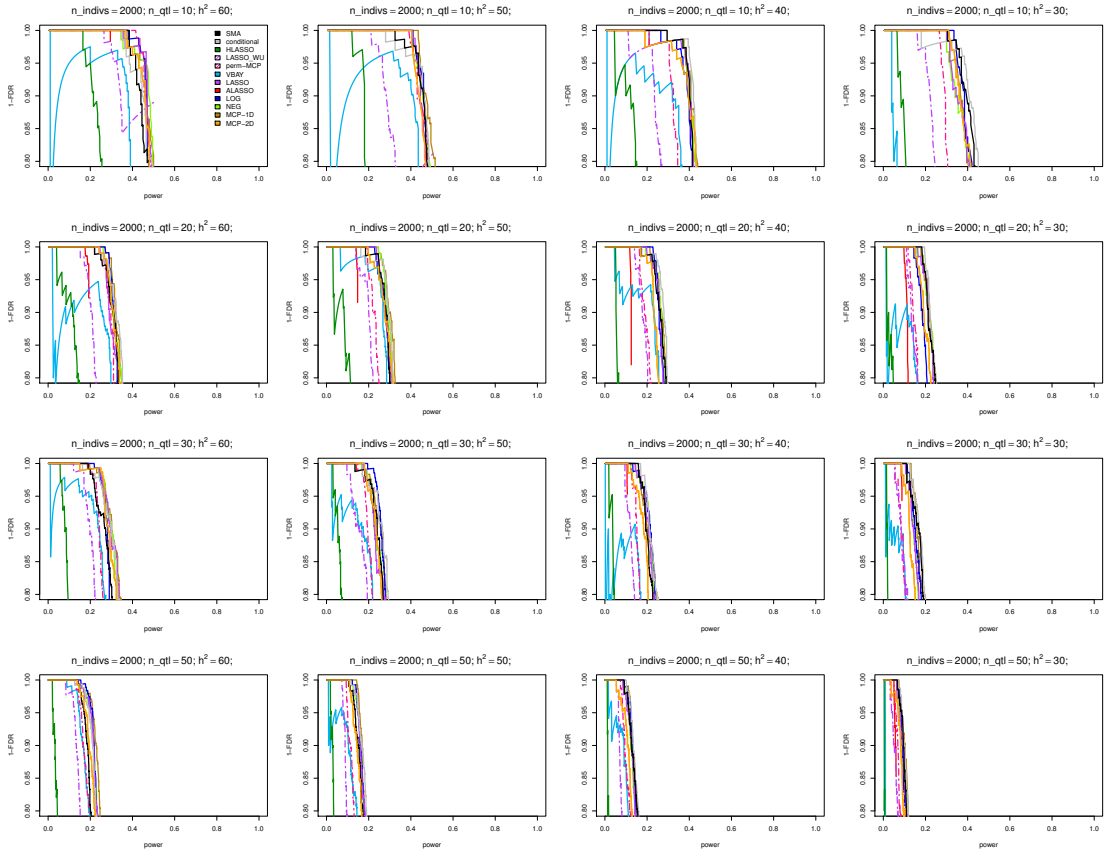


Figure A.10: Simulation results showing precision-recall curves for a sample size of 5000. Results are shown for a range of total heritabilities and number of causal markers. Solid colors for pML methods indicate results using our method for assessing significance in the presence of correlated markers, while dashes indicate the significance method of Wu et al. [209] and perm-MCP [9]. This figures is analogous to Figure 2.3b in the main text.

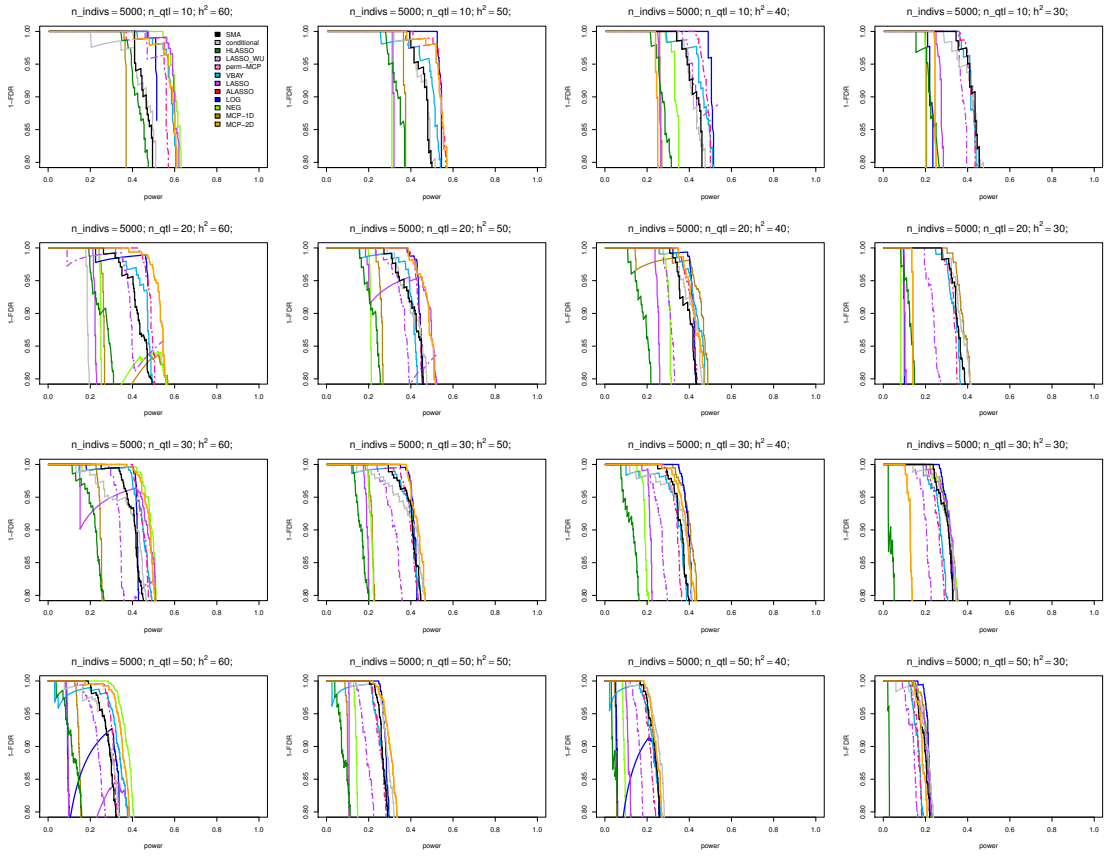


Figure A.11: Precision-Recall curves for perm-MCP for multiple values of eFPR and pre-screening p-value cutoff. Simulations for 5000 samples, 20 causal markers and heritability of 50% using eFPR values $(1 \times 10^{-3}, 1 \times 10^{-4}, 1 \times 10^{-5}, 1 \times 10^{-6}, 1 \times 10^{-7})$ and pre-screening cutoff values (0.1, 0.01, 0.001) indicated in the legend. Results from single marker analysis and MCP-2D are shown for comparison.

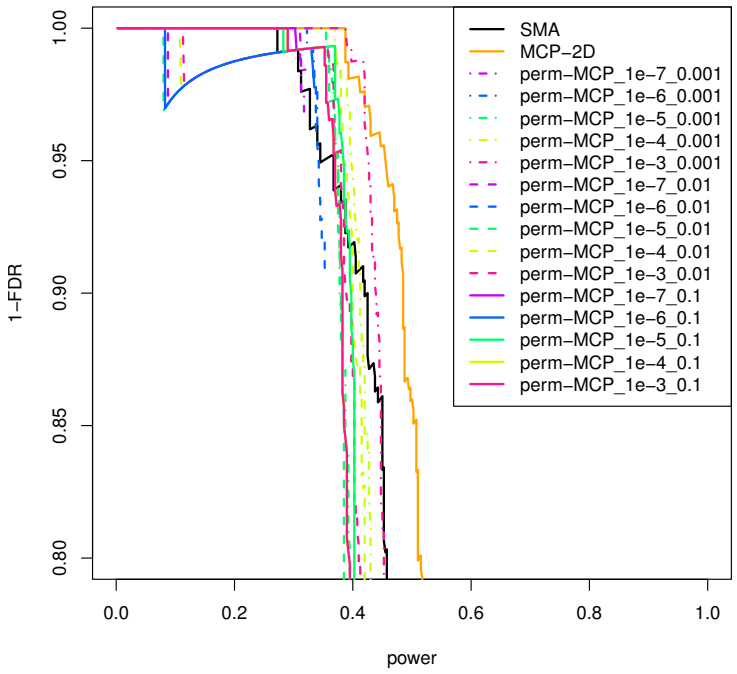
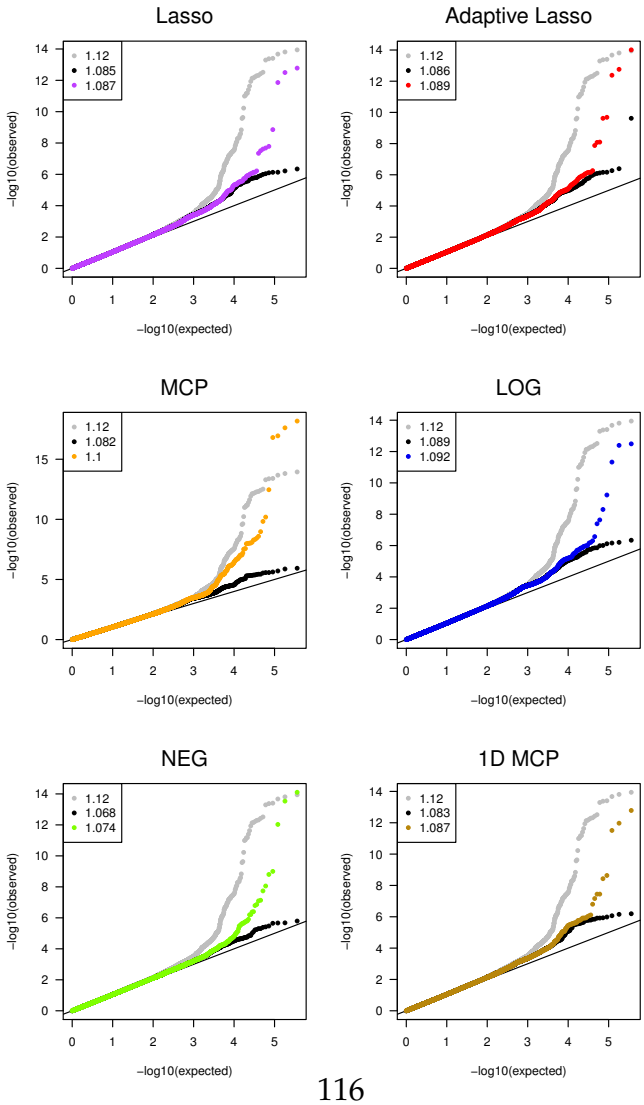
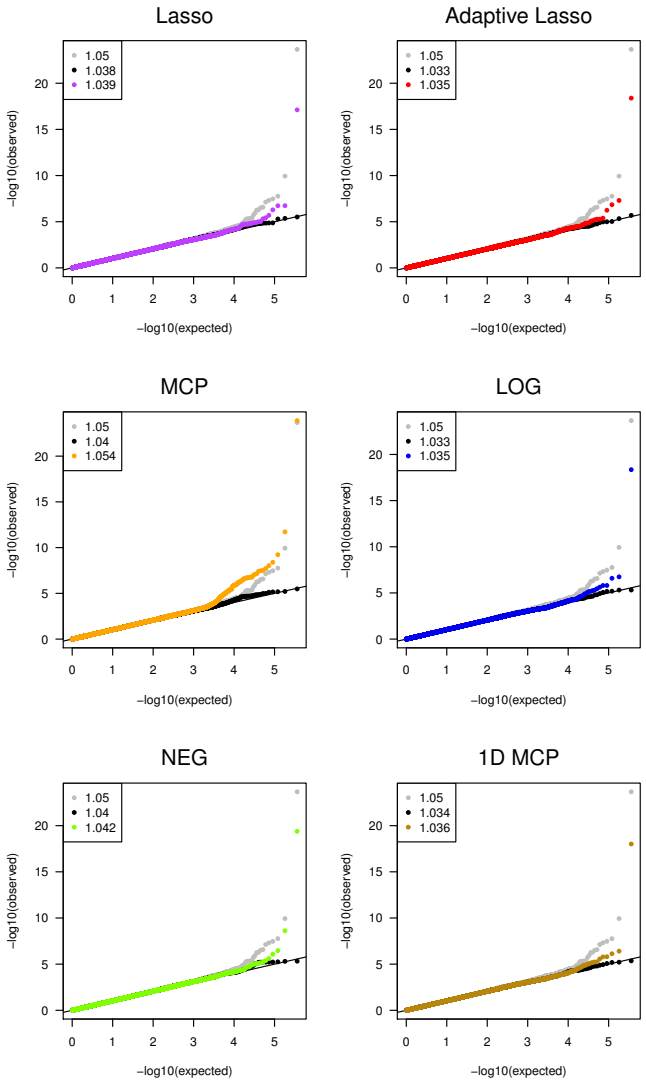


Figure A.12: Quantile-Quantile plots for each disease and method. Plots are shown for a) Crohn's disease, b) Rheumatoid arthritis and c) Type 1 diabetes. Results from a standard single marker analysis of each disease are shown in grey and are the same in all plots for a given disease. Results from including the subset of significantly associated markers identified by each pML method as covariates in a single marker analysis of remaining markers is shown in black, where the relevant method is indicated above each plot. Results from replacing the p-values from this latter analysis with p-values from the PMR method for the relevant markers with nonzero coefficients are shown in color. The genomic inflation values for are shown in the upper left of each plot. Note that the NEG method failed for the type 1 diabetes dataset, so no plot is shown.

(a) Crohn's disease



(b) Rheumatoid arthritis



(c) Type 1 diabetes

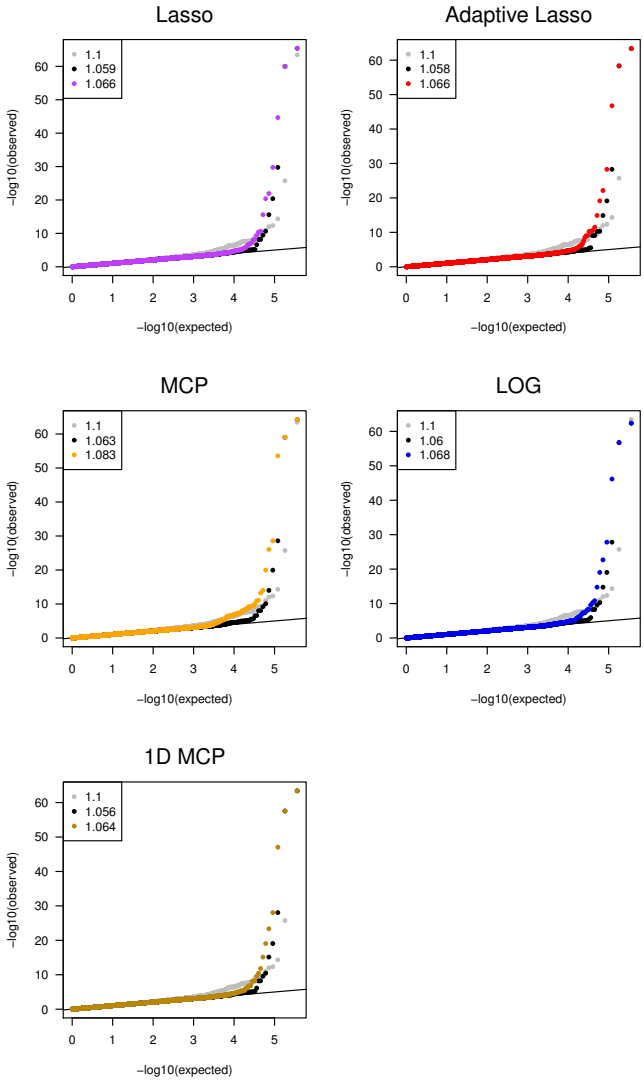
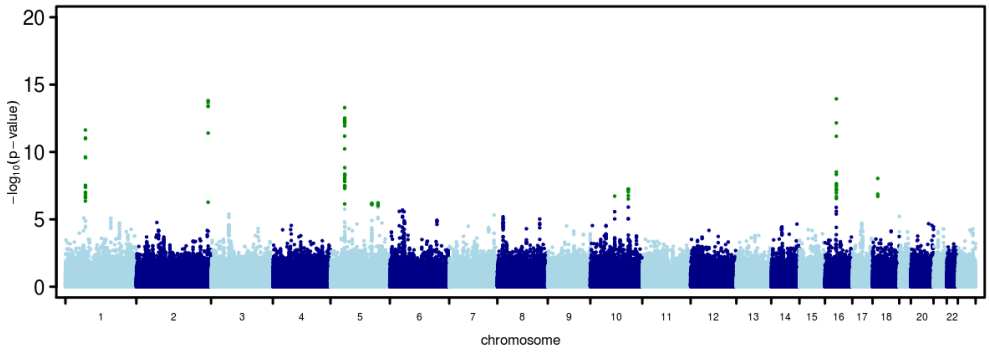
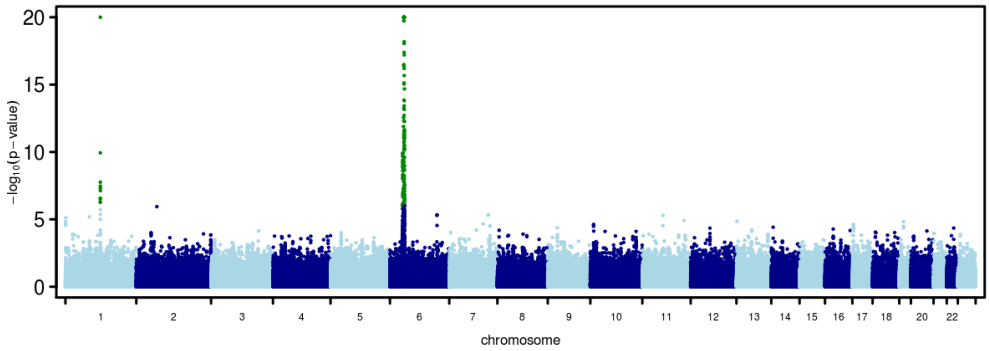


Figure A.13: Manhattan plot showing results of single marker analysis for three disease datasets from our re-analysis. Shown are $-\log_{10}$ p-values where large values are truncated at 20. Markers with $-\log_{10}$ p-values > 6 are colored green.

(a) Crohn's disease



(b) Rheumatoid arthritis



(c) Type 1 diabetes

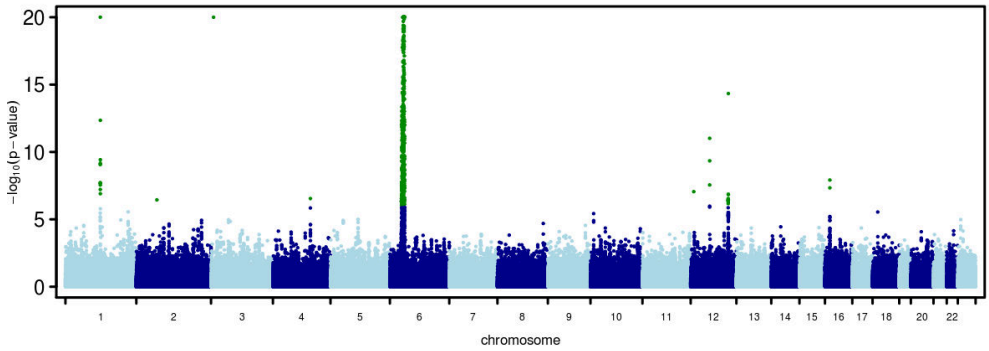
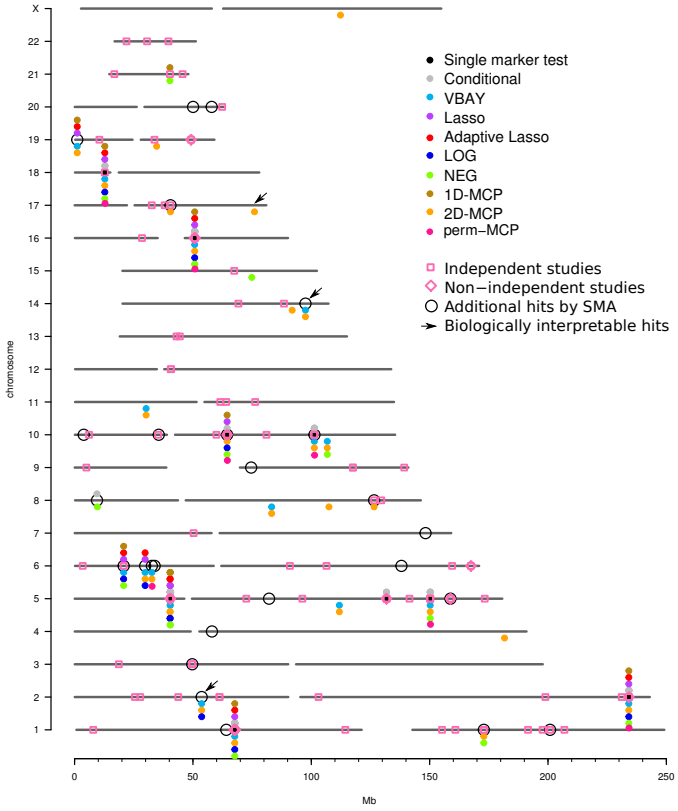
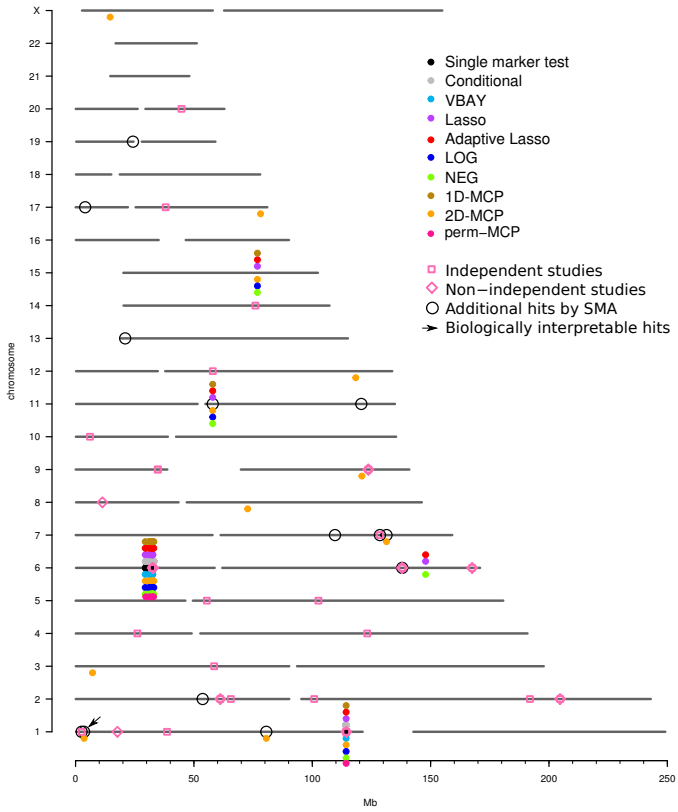


Figure A.14: Genome-wide plot of associations identified by analyzing the WTCCC data for a) Crohn's disease, b) rheumatoid arthritis and c) type 1 diabetes using PMR methods, conditional regression, and single marker analysis. External associations from independent datasets (which do not include WTCCC data) and non-independent datasets (which include WTCCC data) of the same disease are indicated with pink boxes and diamonds, respectively. Markers that are considered associations only when the p-value threshold for the single marker analysis is relaxed to match the same number of associations (with hits in the MHC region excluded) as the union of all PMR methods are indicated with black circles. Arrows indicate novel associations that are biologically interpretable.

(a) Crohn's disease



(b) Rheumatoid arthritis



(c) Type 1 diabetes

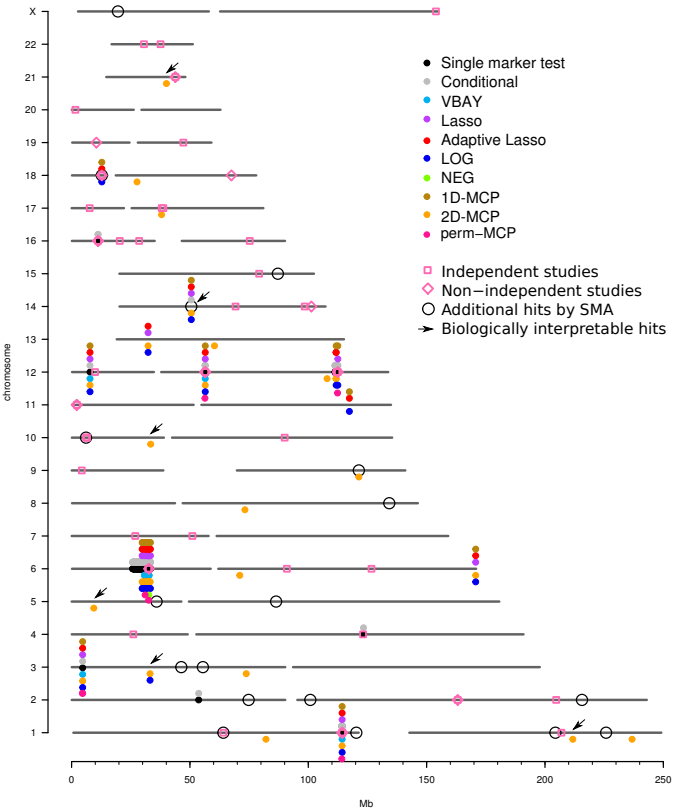
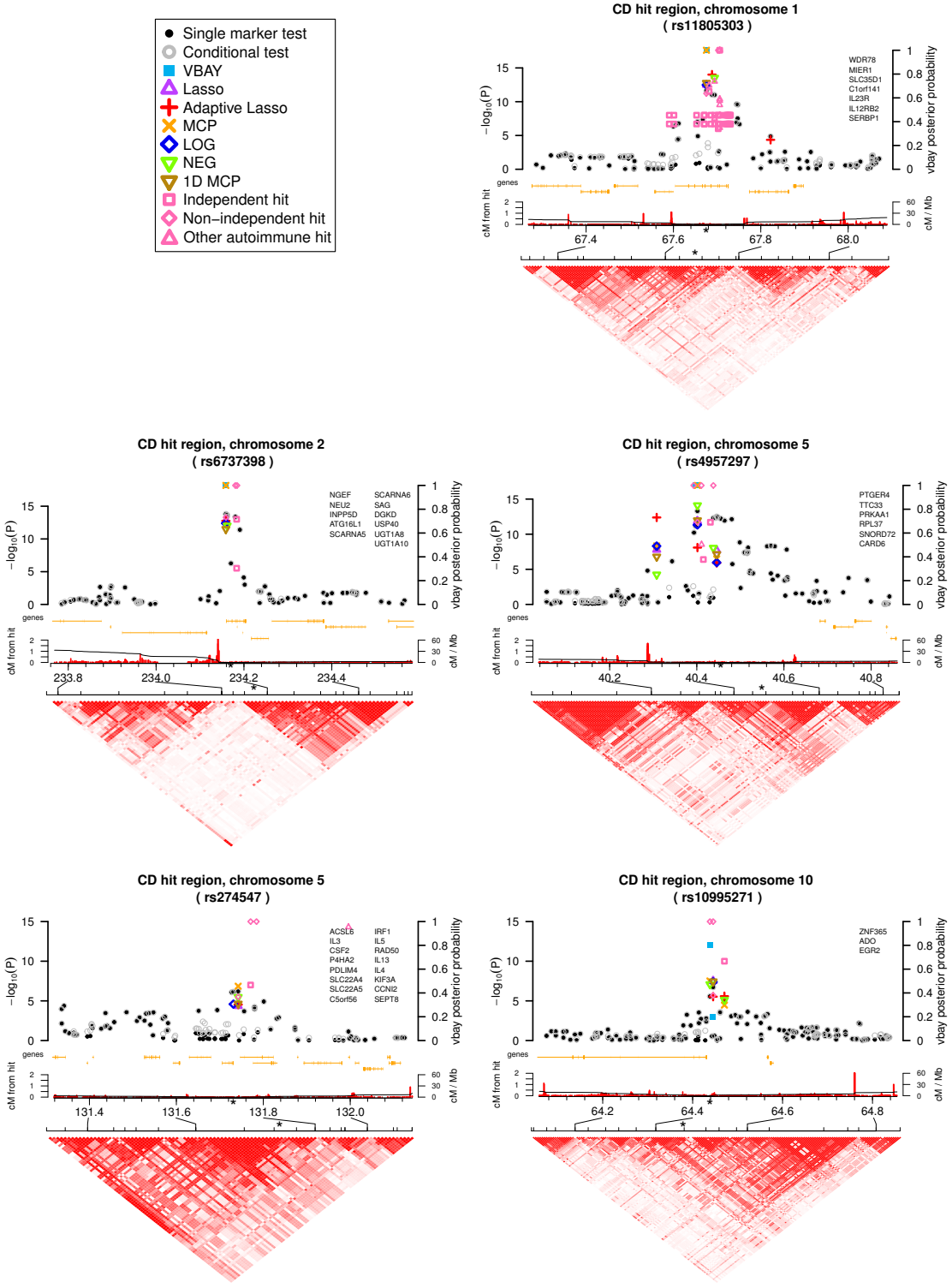


Figure A.15: Local manhattan plots of hits replicated from an independent study of Crohn's disease



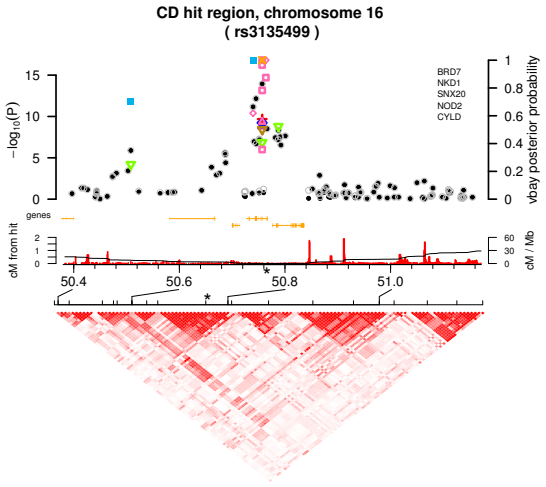

CD hit region, chromosome 16
(rs3135499)
−log10(P)
vbay posterior probability
BRD7
NKD1
SNX20
NOD2
CYLD
genes
cM from hit
cM / Mb
50.4
50.6
50.8
51.0

Figure A.16: Local manhattan plots of hits replicated from an independent study of rheumatoid arthritis

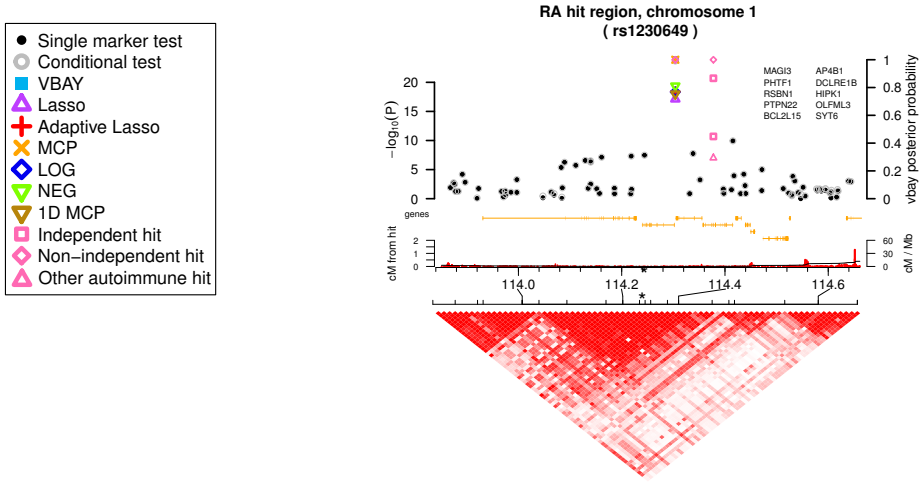


Figure A.17: Local manhattan plots of hits replicated from an independent study of type 1 diabetes

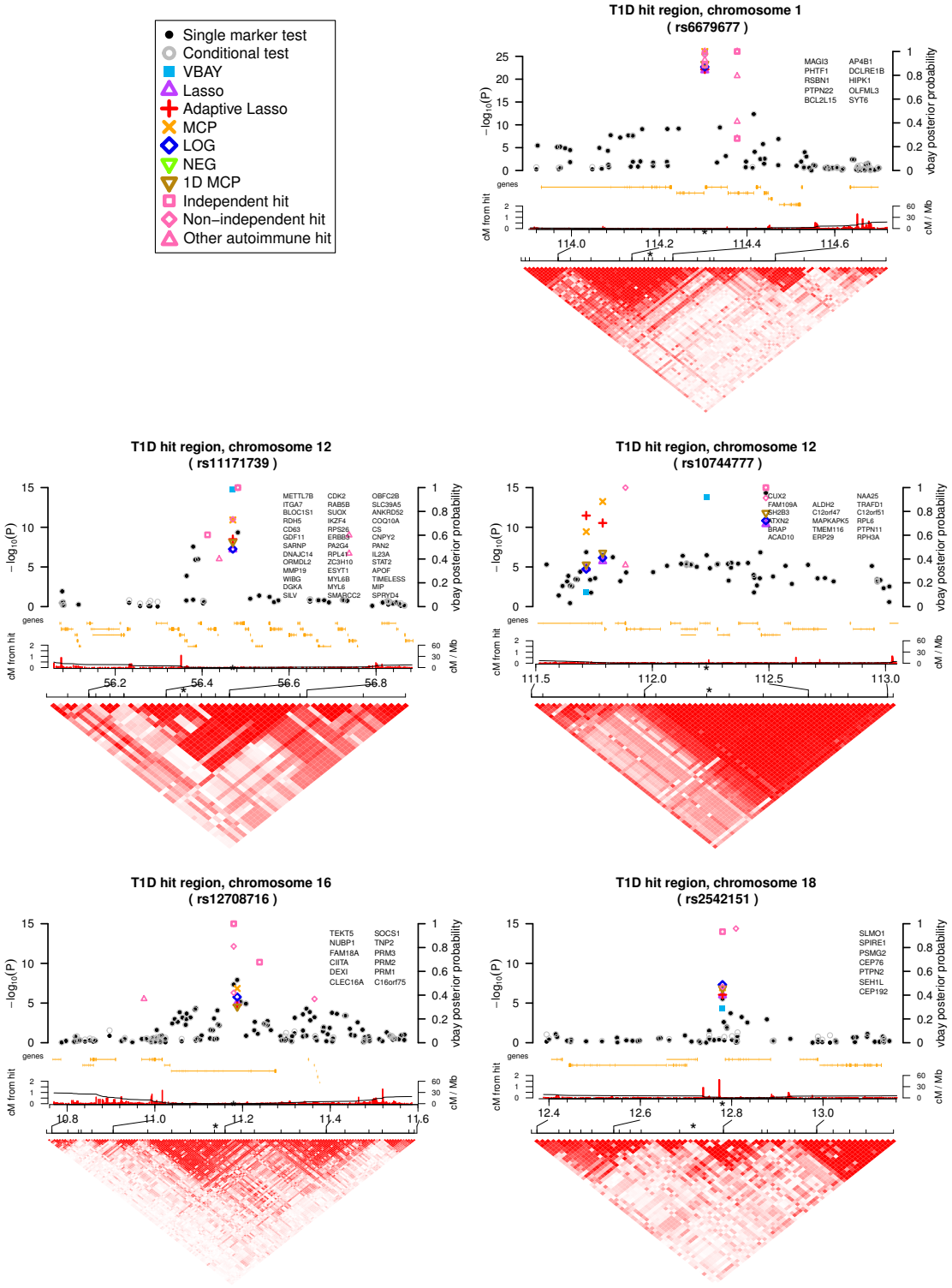
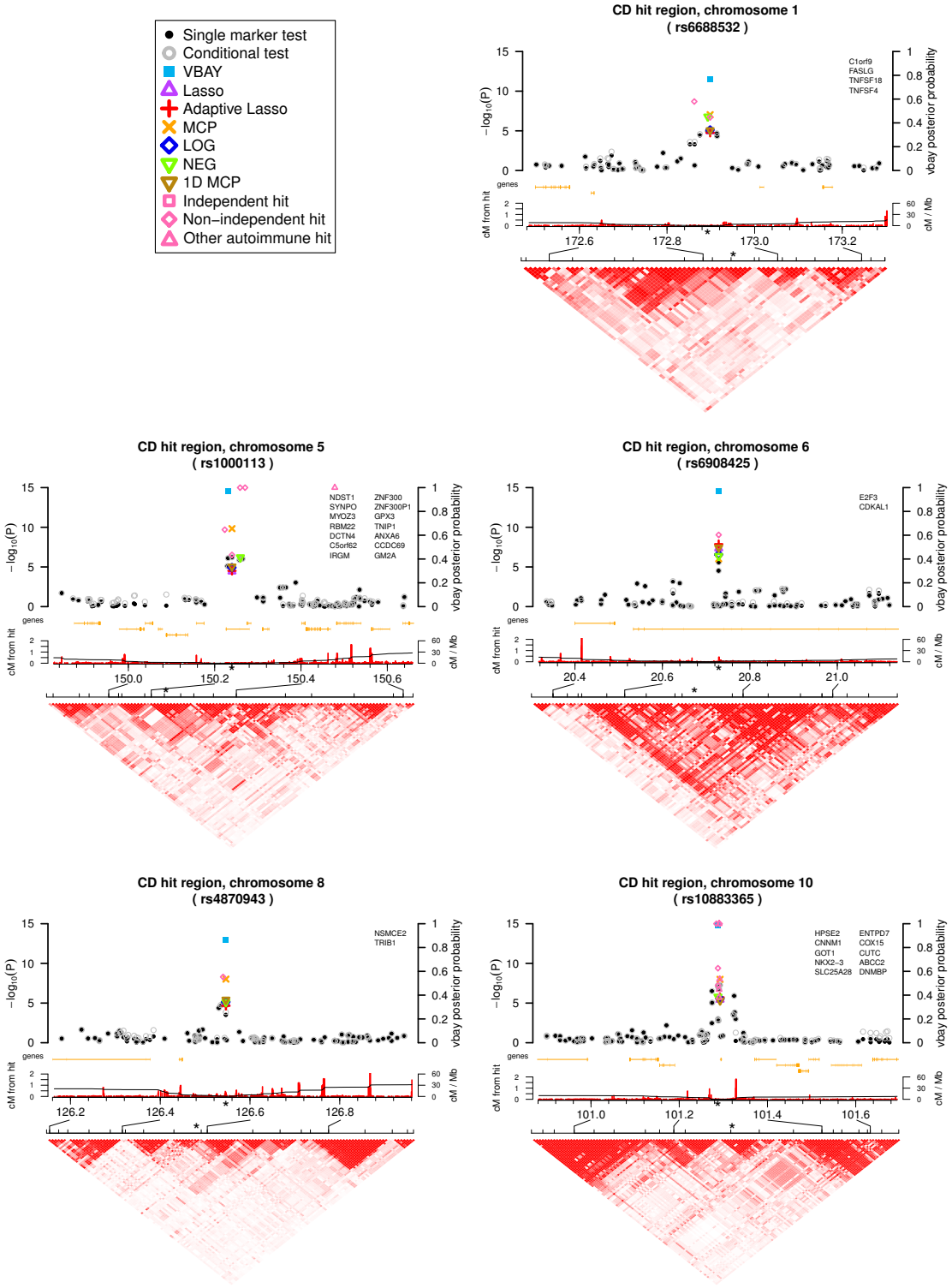
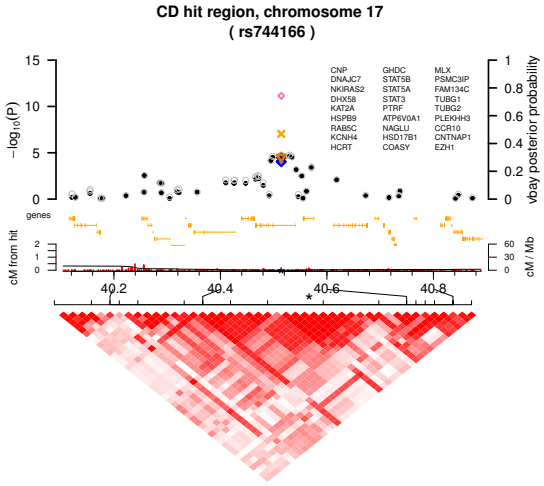
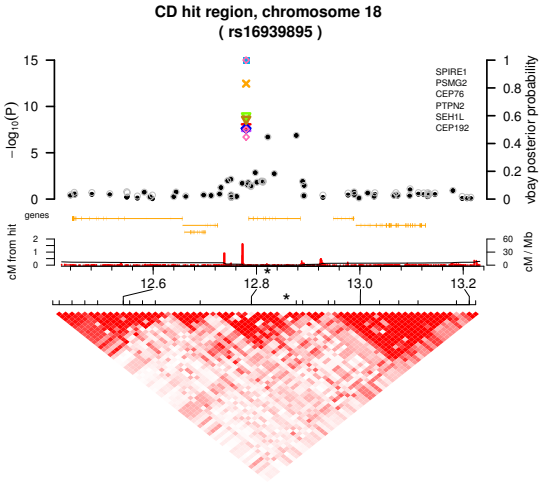


Figure A.18: Local manhattan plots hits replicated from a non-independent study of Crohn's disease




CD hit region, chromosome 17
(rs744166)
CNP
DNAJC7
NKIRAS2
DHX58
KAT2A
HSPB9
RABSC
KCNH4
HCRT
GHDC
STAT5B
STAT5A
STAT3
PTRF
ATP6V0A1
NAGLU
HSD17B1
COASY
MLX
PSMC3IP
FAM134C
TUBG1
TUBG2
PLEKHH3
CCR10
CNTNAP1
EZH1
genes
cM from hit
cM / Mb
vbay posterior probability


CD hit region, chromosome 18
(rs16939895)
SPIRE1
PSMG2
CEP76
PTPN2
SEH1L
CEP192
genes
cM from hit
cM / Mb
vbay posterior probability

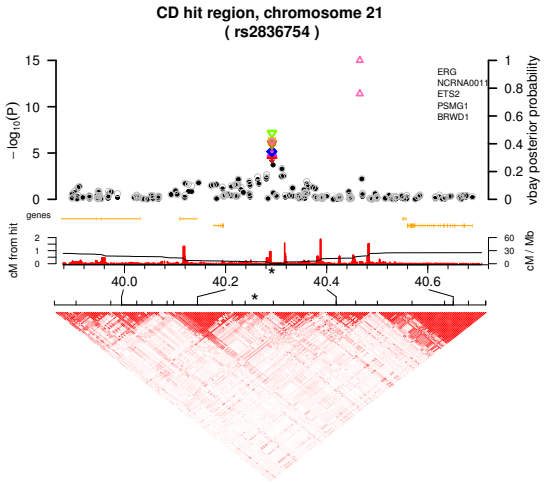

CD hit region, chromosome 21
(rs2836754)
ERG
NCRNA0011
ETS2
PSMG1
BRWD1
genes
cM from hit
cM / Mb
vbay posterior probability

Figure A.19: Local manhattan plots of hits replicated from a non-independent study of type 1 diabetes

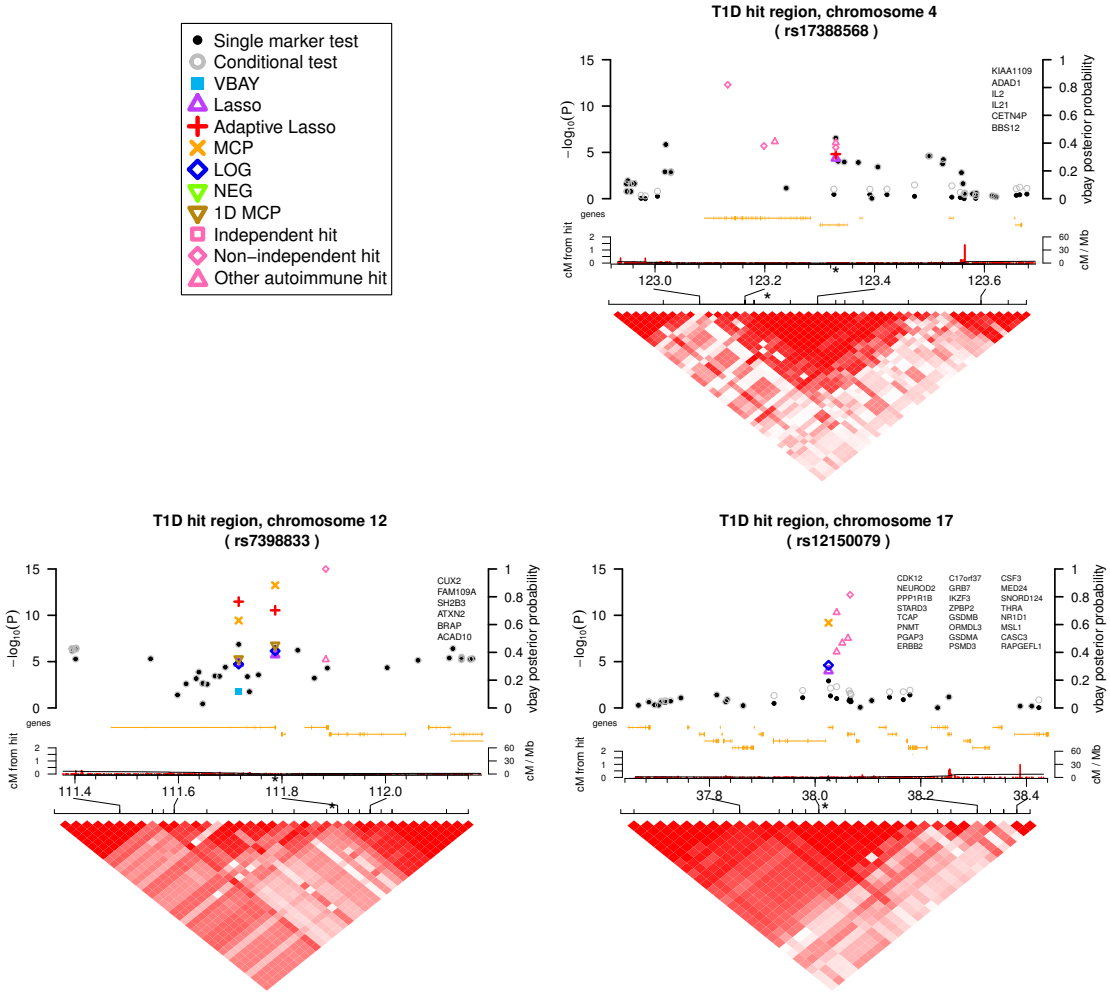


Figure A.20: Local manhattan plots of biologically relevant hits for Crohn's disease

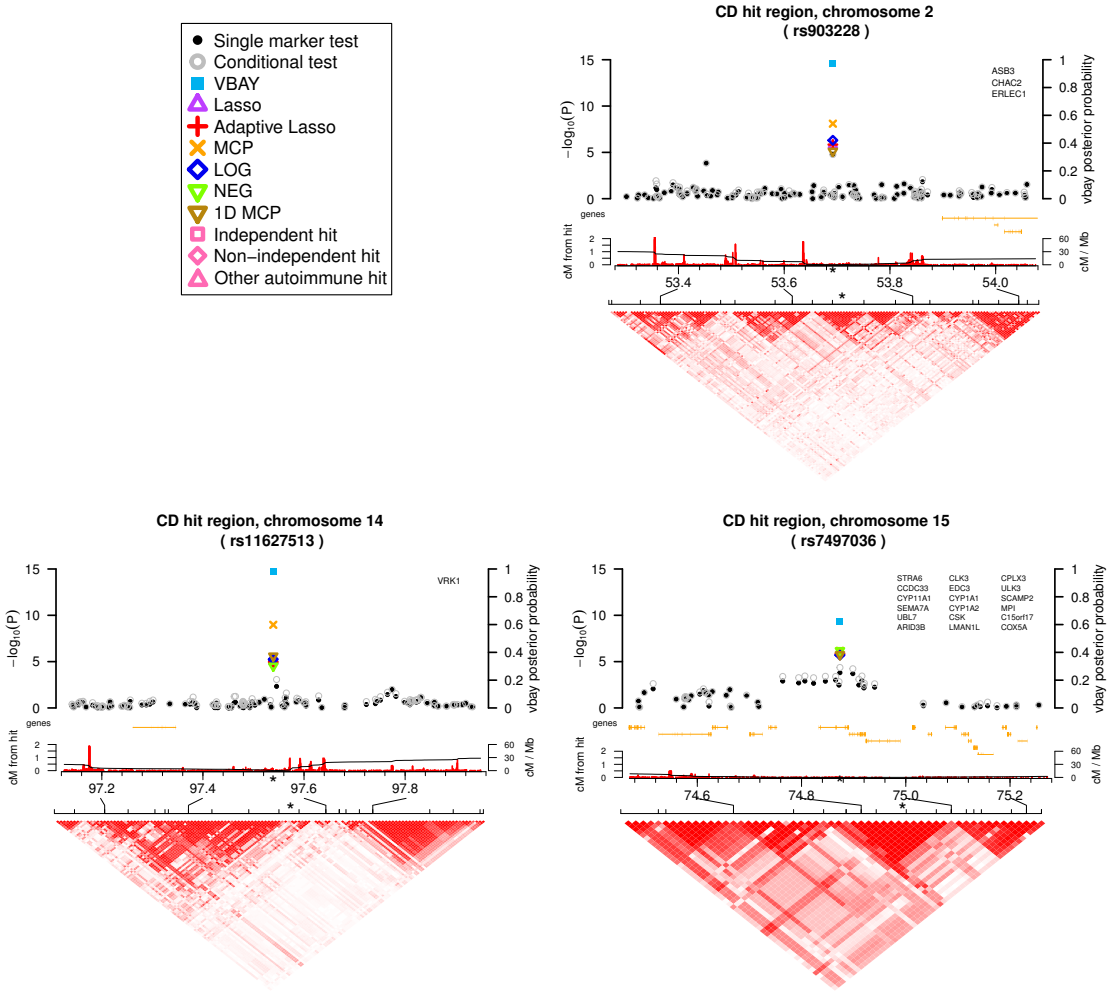


Figure A.21: Local manhattan plots of biologically relevant hits for rheumatoid arthritis

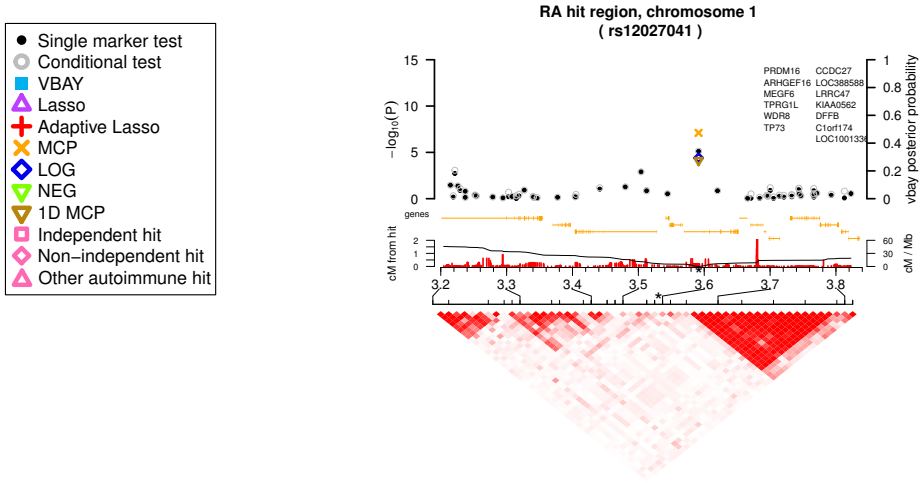
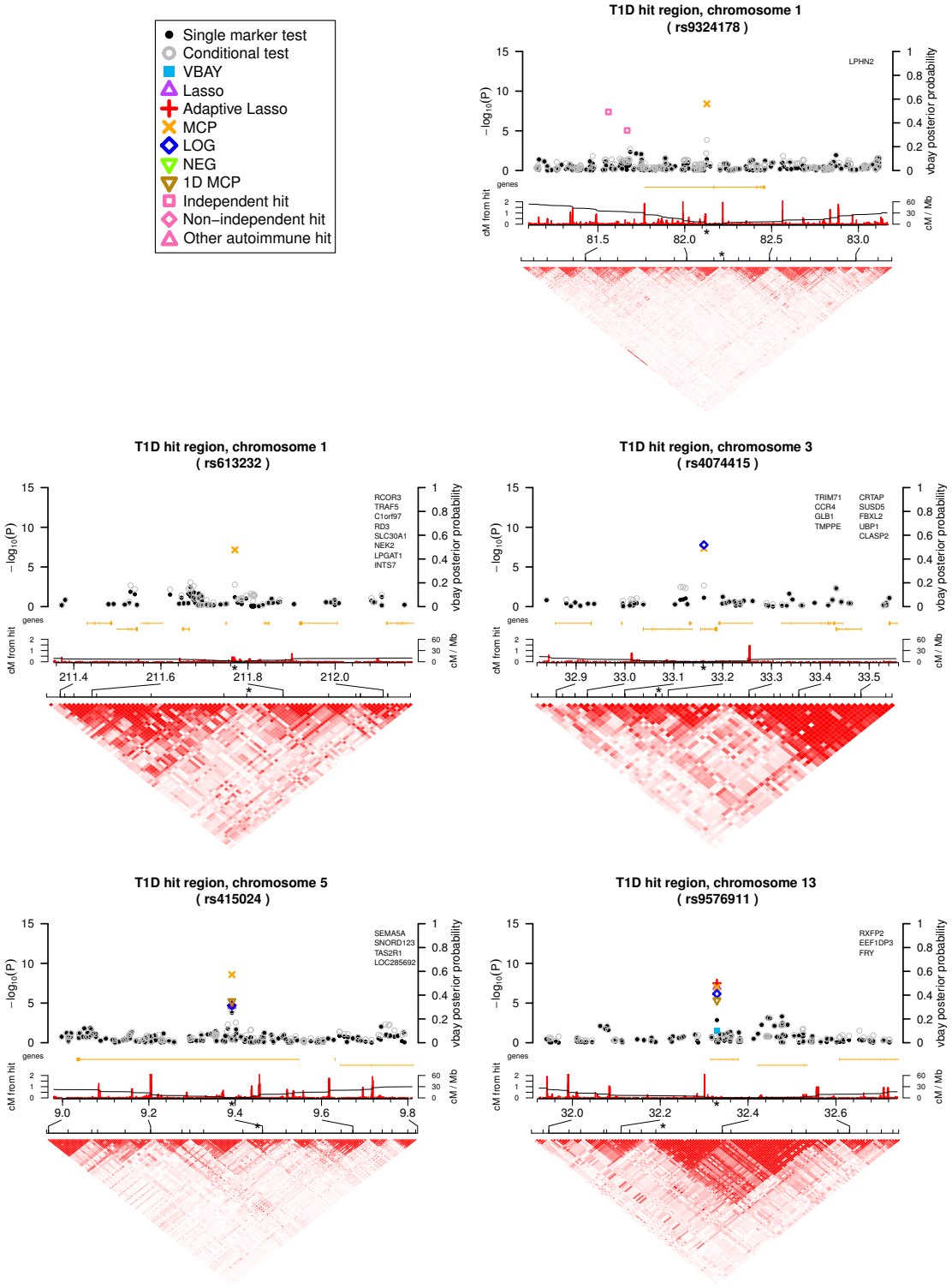
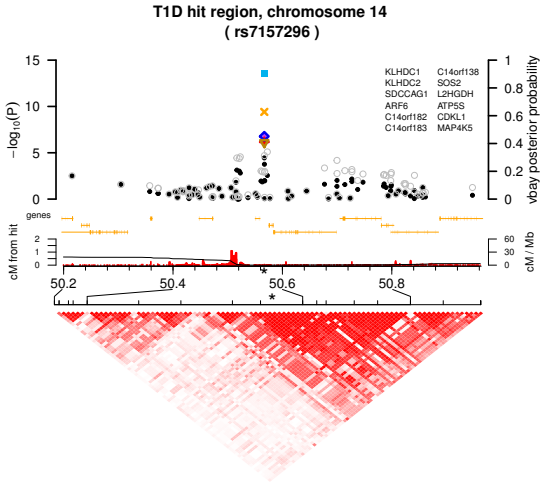


Figure A.22: Local manhattan plots of biologically relevant hits for type 1 diabetes




T1D hit region, chromosome 14
(rs7157296)
KLHDC1
KLHDC2
SDCCAG1
ARF6
C14orf182
C14orf183
C14orf138
SOS2
L2HGDH
ATPSS
CDKL1
MAP4K5
genes
cM from hit
cM / Mb
vbay posterior probability
50.2
50.4
50.6
50.8

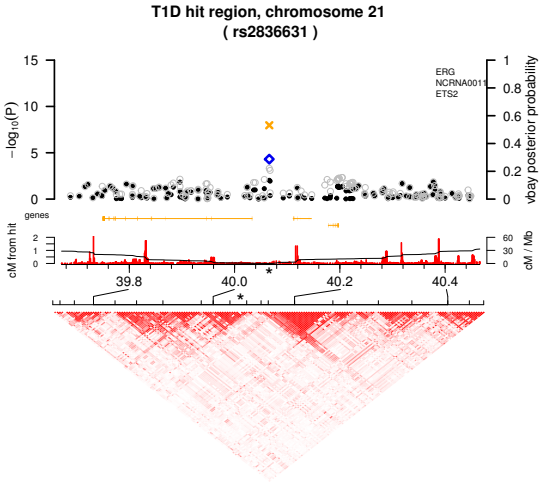

T1D hit region, chromosome 21
(rs2836631)
ERG
NCRNA0011
ETS2
genes
cM from hit
cM / Mb
vbay posterior probability
39.8
40.0
40.2
40.4

Table A.1: Concordance of PMR hits with single marker analysis. Number of regions identified by a single marker analysis with a p-value $< 1 \times 10^{-6}$ and the number of these regions that are recapitulated by each other method.

	Method									
disease	SMA	Conditional	VBAY	Lasso	Adaptive Lasso	2D-MCP	LOG	NEG	1D-MCP	perm-MCP
CD	9	9	7	6	5	8	6	7	6	6
RA	1	1	1	1	1	1	1	1	1	1
T1D	9	9	3	4	4	4	4	0	4	4
Total	19	19	11	11	10	13	11	8	11	11

Table A.2: Regions identified by single marker analysis with a p-value $< 1 \times 10^{-6}$. Of the regions missed by PMR methods, only one association passes the Bonferroni cutoff of $0.05/360,000 = 1.38 \times 10^{-7}$, although 5 of these have been replicated: rs6596075 [11, 46, 126], rs17388568 [10], rs10807124 [10, 30, 59], rs4766517 [10, 30, 59], rs12924729 [10, 30, 59, 190].

				Method									
disease	SNP	chromosome	position	SMA	Conditional	VBAY	Lasso	Adaptive Lasso	2D-MCP	LOG	NEG	1D-MCP	perm-MCP
CD	rs11805303	1p31.3	67,675,515	2.34×10^{-12}	2.34×10^{-12}	1	3.18×10^{-13}	9.77×10^{-15}	2.43×10^{-18}	3.23×10^{-13}	2.92×10^{-14}	1.64×10^{-13}	-
CD	rs10210302	2q37.1	234,158,838	1.54×10^{-14}	1.54×10^{-14}	1	1.66×10^{-13}	1.73×10^{-13}	6.69×10^{-19}	4.04×10^{-13}	9.32×10^{-13}	3.09×10^{-12}	1.19×10^{-13}
CD	rs17234657	5p13.1	40,401,508	5.09×10^{-14}	5.09×10^{-14}	1	1.38×10^{-12}	4.15×10^{-13}	1.13×10^{-17}	4.71×10^{-12}	7.84×10^{-15}	1.07×10^{-12}	6.10×10^{-06}
CD	rs6596075	5q31.1	131,742,227	6.37×10^{-07}	6.37×10^{-07}	0.064	4.4×10^{-05}	2.95×10^{-05}	1.42×10^{-07}	2.58×10^{-05}	3.29×10^{-06}	2.3×10^{-05}	-
CD	rs1000113	5q33.1	150,240,075	5.99×10^{-07}	5.99×10^{-07}	0.973	3.69×10^{-05}	3.11×10^{-05}	1.49×10^{-10}	1.21×10^{-05}	6.14×10^{-07}	9.02×10^{-06}	5.04×10^{-07}
CD	rs10761659	10q21.2	64,445,563	1.9×10^{-07}	1.9×10^{-07}	0.8	3.13×10^{-08}	2.96×10^{-06}	3.11×10^{-08}	4.07×10^{-08}	8.03×10^{-08}	3.58×10^{-08}	7.41×10^{-08}
CD	rs10883371	10q24.2	101,292,454	5.55×10^{-08}	5.55×10^{-08}	0.99	3.78×10^{-06}	1.87×10^{-06}	1.03×10^{-08}	2.85×10^{-06}	1.57×10^{-06}	4.83×10^{-06}	2.84×10^{-07}
CD	rs2076756	16q12.1	50,756,880	1.14×10^{-14}	1.14×10^{-14}	1	1.4×10^{-09}	2.06×10^{-10}	1.56×10^{-17}	5.95×10^{-10}	1.58×10^{-09}	3.72×10^{-09}	8.93×10^{-09}
CD	rs2542151	18p11.21	12,779,946	9.2×10^{-09}	9.2×10^{-09}	1	2.44×10^{-08}	8.41×10^{-09}	3.45×10^{-13}	2.32×10^{-08}	1×10^{-09}	2.32×10^{-09}	7.56×10^{-08}
RA	rs6679677	1p13.2	114,303,807	2.16×10^{-24}	2.16×10^{-24}	1	7.64×10^{-18}	4.1×10^{-19}	1.31×10^{-24}	4.48×10^{-19}	4.11×10^{-20}	9.63×10^{-19}	7.53×10^{-20}
T1D	rs6679677	1p13.2	114,303,807	1.8×10^{-26}	1.8×10^{-26}	1	1.17×10^{-22}	7.17×10^{-23}	8.74×10^{-27}	2.04×10^{-23}	-	4.35×10^{-24}	2.25×10^{-20}
T1D	rs903228	2p16.2	53,692,048	3.59×10^{-07}	3.59×10^{-07}	0.000 398	-	-	-	-	-	-	-
T1D	rs17388568	4q27	123,329,361	2.87×10^{-07}	2.87×10^{-07}	0.002 93	4.33×10^{-05}	1.54×10^{-05}	-	-	-	-	-
T1D	rs10807124	6p21.32	33,404,063	2.07×10^{-09}	2.07×10^{-09}	7.02×10^{-05}	-	-	-	-	-	-	-
T1D	rs2110221	12p13.31	7,794,773	8.75×10^{-08}	8.75×10^{-08}	0.987	2.61×10^{-08}	2.19×10^{-08}	2.44×10^{-09}	4.49×10^{-08}	-	1.12×10^{-08}	-
T1D	rs11171739	12q13.2	56,470,624	9.67×10^{-12}	9.67×10^{-12}	0.983	6.31×10^{-08}	3.18×10^{-09}	1.2×10^{-11}	5.52×10^{-08}	-	7.2×10^{-09}	6.25×10^{-07}
T1D	rs4766517	12q24.11	111,359,711	2.87×10^{-07}	2.87×10^{-07}	0.000 433	-	-	-	-	-	-	-
T1D	rs17696736	12q24.13	112,486,817	4.62×10^{-15}	4.62×10^{-15}	0.923	4.04×10^{-11}	2.89×10^{-11}	5.74×10^{-14}	1.47×10^{-11}	-	1.54×10^{-12}	2.66×10^{-07}
T1D	rs12924729	16p13.13	11,187,782	1.21×10^{-08}	1.21×10^{-08}	0.027	1.05×10^{-05}	1.77×10^{-05}	1.46×10^{-07}	1.72×10^{-06}	-	2.8×10^{-05}	-

Table A.3: Regions which are significant either by single marker analysis, conditional regression, or a PMR method and which recapitulate a known association to the same disease in an independent study that does not include data from the WTCCC. The table includes all regions with a VBAY posterior probability > 0.97, an MCP p-value $< 1 \times 10^{-7}$, or a p-value for any other method $< 1 \times 10^{-6}$.

				Method											
disease	SNP	chromosome	position	SMA	Conditional	VBAY	Lasso	Adaptive Lasso	2D-MCP	LOG	NEG	1D-MCP	perm-MCP	genes	references
CD	rs11805303	1p31.3	67,675,515	2.34×10^{-12}	2.34×10^{-12}	1	3.18×10^{-13}	9.77×10^{-15}	2.43×10^{-18}	3.23×10^{-13}	2.92×10^{-14}	1.64×10^{-13}	-	IL23R	[106, 126, 155, 158]
CD	rs6737398	2q37.1	234,170,396	1.54×10^{-14}	1.54×10^{-14}	1	1.66×10^{-13}	1.73×10^{-13}	6.69×10^{-19}	4.04×10^{-13}	9.32×10^{-13}	3.09×10^{-12}	1.19×10^{-13}	ATG16L1	[126, 158]
CD	rs4957297	5p13.1	40,455,073	5.09×10^{-14}	5.09×10^{-14}	1	1.38×10^{-12}	8.02×10^{-09}	1.13×10^{-17}	4.71×10^{-12}	7.84×10^{-15}	1.07×10^{-12}	-	Intergenic, PTGER4	[45, 106]
CD	rs274547	5q31.1	131,731,303	6.37×10^{-07}	6.37×10^{-07}	0.064	4.4×10^{-05}	2.95×10^{-05}	1.42×10^{-07}	2.58×10^{-05}	3.29×10^{-06}	2.3×10^{-05}	-	IBD5	[126]
CD	rs10995271	10q21.2	64,438,485	1.9×10^{-07}	1.9×10^{-07}	0.8	3.13×10^{-08}	2.96×10^{-06}	3.11×10^{-08}	4.07×10^{-08}	8.03×10^{-08}	3.58×10^{-08}	-	Intergenic	[158]
CD	rs3135499	16q12.1	50,766,126	1.14×10^{-14}	1.14×10^{-14}	1	1.4×10^{-09}	2.06×10^{-10}	1.56×10^{-17}	5.95×10^{-10}	1.58×10^{-09}	3.72×10^{-09}	-	NOD2	[45, 106, 126, 155, 158]
RA	rs1230649	1p13.2	114,244,176	2.16×10^{-24}	2.16×10^{-24}	1	7.64×10^{-18}	4.1×10^{-19}	1.31×10^{-24}	4.48×10^{-19}	4.11×10^{-20}	9.63×10^{-19}	-	PTPN22	[53, 147]
T1D	rs6679677	1p13.2	114,303,807	1.8×10^{-26}	1.8×10^{-26}	1	1.17×10^{-22}	7.17×10^{-23}	8.74×10^{-27}	2.04×10^{-23}	-	4.35×10^{-24}	-	PHTF1, PTPN22	[59, 190]
T1D	rs11171739	12q13.2	56,470,624	9.67×10^{-12}	9.67×10^{-12}	0.983	6.31×10^{-08}	3.18×10^{-09}	1.2×10^{-11}	5.52×10^{-08}	-	7.2×10^{-09}	-	ERBB3, RAB5B, SUOX, IKZF4, CDK	[60, 190]
T1D	rs10744777	12q24.12	112,233,017	3.34×10^{-07}	3.34×10^{-07}	0.923	5.28×10^{-01}	-	-	-	-	1.17×10^{-01}	-	C12orf30	[190]
T1D	rs12708716	16p13.13	11,179,872	1.21×10^{-08}	1.21×10^{-08}	0.027	1.05×10^{-05}	1.77×10^{-05}	1.46×10^{-07}	1.72×10^{-06}	-	2.8×10^{-05}	-	CLEC16A	[59, 190]
T1D	rs2542151	18p11.21	12,779,946	2.86×10^{-06}	2.86×10^{-06}	0.288	1.06×10^{-06}	9×10^{-07}	1.15×10^{-07}	4.89×10^{-08}	-	1.43×10^{-07}	-	PTPN2	[190]

Table A.4: Regions which are significant either by single marker analysis, conditional regression, or a PMR method and which recapitulate a known association to the same disease in a non-independent study that includes data from the WTCCC. The table includes all regions with a VBAY posterior probability > 0.97, an MCP p-value $< 1 \times 10^{-7}$, or a p-value for any other method $< 1 \times 10^{-6}$.

				Method											
disease	SNP	chromosome	position	SMA	Conditional	VBAY	Lasso	Adaptive Lasso	2D-MCP	LOG	NEG	1D-MCP	perm-MCP	genes	references
CD	rs6688532	1q24.3	172,892,951	8.35×10^{-06}	5.53×10^{-06}	0.764	1.15×10^{-05}	1.74×10^{-05}	9.26×10^{-08}	7.83×10^{-06}	1.61×10^{-07}	8.56×10^{-06}	-	Intergenic	[11, 142]
CD	rs1000113	5q33.1	150,240,075	5.99×10^{-07}	5.99×10^{-07}	0.973	3.69×10^{-05}	3.11×10^{-05}	1.49×10^{-10}	1.21×10^{-05}	6.14×10^{-07}	9.02×10^{-06}	-	IRGM	[11, 46, 142, 205]
CD	rs6908425	6p22.3	20,728,730	2.57×10^{-06}	2.57×10^{-06}	0.973	4.54×10^{-08}	1.32×10^{-08}	9.3×10^{-07}	2.69×10^{-07}	4.18×10^{-07}	3.59×10^{-08}	-	CDKAL1	[11]
CD	rs4870943	8q24.13	126,546,388	2.01×10^{-05}	2.01×10^{-05}	0.864	1.14×10^{-05}	2.24×10^{-05}	9.59×10^{-09}	7.12×10^{-06}	7.73×10^{-06}	3.78×10^{-06}	-	Intergenic	[11]
CD	rs10883365	10q24.2	101,287,763	5.55×10^{-08}	5.55×10^{-08}	0.99	3.78×10^{-06}	1.87×10^{-06}	1.03×10^{-08}	2.85×10^{-06}	1.57×10^{-06}	4.83×10^{-06}	-	NKX2-3	[11, 46, 142, 205]
CD	rs744166	17q21.2	40,514,200	1.9×10^{-05}	1.78×10^{-05}	0.0246	5.53×10^{-05}	3.58×10^{-05}	9.21×10^{-08}	9.31×10^{-05}	1.84×10^{-01}	2.97×10^{-05}	-	STAT3	[11]
CD	rs16939895	18p11.21	12,821,902	9.2×10^{-09}	9.2×10^{-09}	1	2.44×10^{-08}	8.41×10^{-09}	3.45×10^{-13}	2.32×10^{-08}	1×10^{-09}	2.32×10^{-09}	-	PTPN2	[11, 142, 205]
CD	rs2836754	21q22.2	40,291,739	5.22×10^{-05}	2.66×10^{-05}	0.0487	1.19×10^{-05}	3.31×10^{-05}	5.88×10^{-07}	7×10^{-06}	7.34×10^{-08}	9.66×10^{-07}	-	Intergenic	[142]
T1D	rs17388568	4q27	123,329,361	2.87×10^{-07}	2.87×10^{-07}	-	4.33×10^{-05}	1.54×10^{-05}	-	-	-	-	-	IL2	[10]
T1D	rs7398833	12q24.12	111,786,891	1.36×10^{-07}	1.36×10^{-07}	0.119	1.89×10^{-06}	3.34×10^{-12}	5.74×10^{-14}	7.09×10^{-07}	-	1.63×10^{-07}	-	C12orf30, SH2B3	[10, 30]
T1D	rs12150079	17q12	38,025,416	1.21×10^{-03}	2.83×10^{-05}	0.0153	1×10^{-04}	1.8×10^{-04}	6.32×10^{-10}	2.47×10^{-05}	-	1.91×10^{-04}	-	ORMDL3	[10]

Table A.5: For each method, the number of associations in this re-analysis that replicate associations identified in a) independent (not including WTCC data) and b) non-independent (including WTCC data) datasets, where the number of markers considered as 'hits' is set to be equal across methods. For each method the number of hits is set to a given value and the number of replications is reported. Numbers in parentheses indicate the number of hits that are distinct from those found by the single marker analysis.

a) Independent studies

	# considered as hits	Method								
		SMA	Conditional test	VBAY	Lasso	Adaptive Lasso	2D-MCP	LOG	NEG	1D-MCP
CD	5	4	4 (0)	4 (0)	4 (0)	4 (0)	4 (0)	4 (0)	4 (0)	4 (0)
	10	6	6 (0)	4 (0)	5 (0)	4 (0)	4 (0)	5 (0)	5 (0)	5 (0)
	15	6	6 (0)	4 (0)	5 (0)	5 (0)	4 (0)	5 (0)	5 (0)	5 (0)
	25	7	6 (0)	5 (0)	5 (0)	5 (0)	5 (0)	6 (0)	6 (0)	6 (0)
	30	7	6 (0)	5 (0)	5 (0)	6 (0)	6 (0)	6 (0)	6 (0)	6 (0)
RA	5	2	2 (0)	1 (0)	1 (0)	1 (0)	1 (0)	1 (0)	1 (0)	1 (0)
	10	2	2 (0)	1 (0)	1 (0)	1 (0)	1 (0)	1 (0)	1 (0)	1 (0)
	15	2	2 (0)	1 (0)	1 (0)	1 (0)	1 (0)	1 (0)	1 (0)	1 (0)
	25	2	2 (0)	1 (0)	1 (0)	1 (0)	1 (0)	1 (0)	1 (0)	1 (0)
	30	2	2 (0)	1 (0)	1 (0)	1 (0)	1 (0)	1 (0)	1 (0)	1 (0)
T1D	5	3	3 (0)	3 (0)	3 (0)	2 (0)	3 (0)	2 (0)	0	3 (0)
	10	4	4 (0)	4 (1)	4 (1)	3 (1)	3 (0)	4 (1)	0	4 (1)
	15	5	4 (0)	4 (0)	4 (0)	3 (0)	3 (0)	5 (0)	0	4 (0)
	25	5	5 (0)	5 (0)	5 (0)	4 (0)	4 (0)	5 (0)	0	5 (0)
	30	5	5 (0)	5 (0)	5 (0)	4 (0)	5 (0)	5 (0)	0	5 (0)

b) Non-independent studies

	# considered as hits	Method								
		SMA	Conditional test	VBAY	Lasso	Adaptive Lasso	2D-MCP	LOG	NEG	1-MCP
CD	5	5	5 (0)	5 (0)	5 (0)	5 (0)	5 (0)	5 (0)	5 (0)	5 (0)
	10	9	9 (0)	6 (0)	7 (1)	7 (1)	6 (0)	7 (1)	8 (2)	8 (2)
	15	11	11 (0)	8 (0)	8 (0)	8 (0)	7 (1)	8 (0)	11 (2)	11 (4)
	25	15	13 (0)	12 (2)	13 (3)	12 (2)	11 (1)	15 (3)	14 (2)	16 (4)
	30	16	14 (0)	14 (3)	14 (3)	16 (3)	12 (0)	16 (4)	16 (4)	17 (4)
RA	5	3	3 (1)	2 (1)	1 (0)	1 (0)	1 (0)	1 (0)	1 (0)	1 (0)
	10	3	4 (1)	2 (1)	2 (1)	2 (1)	1 (0)	1 (0)	2 (1)	2 (1)
	15	4	4 (1)	2 (1)	2 (0)	3 (1)	1 (0)	3 (1)	2 (0)	3 (1)
	25	6	5 (0)	3 (0)	3 (0)	3 (0)	2 (0)	3 (0)	4 (0)	3 (0)
	30	6	5 (0)	3 (0)	4 (0)	4 (0)	2 (0)	3 (0)	4 (0)	3 (0)
T1D	5	3	3 (0)	3 (0)	3 (0)	2 (0)	3 (0)	2 (0)	0	3 (0)
	10	4	4 (0)	4 (1)	4 (1)	3 (1)	4 (1)	4 (1)	0	4 (1)
	15	6	5 (0)	5 (1)	4 (0)	3 (0)	4 (1)	5 (0)	0	4 (0)
	25	7	7 (0)	7 (1)	6 (0)	7 (2)	5 (1)	6 (1)	0	5 (0)
	30	8	7 (0)	8 (1)	6 (0)	7 (2)	6 (1)	7 (2)	0	6 (1)

Table A.6: Additional associations for Crohn's disease identified by PMR methods but not a single marker analysis

				Method										
disease	SNP	chromosome	position	SMA	Conditional	VBAY	Lasso	Adaptive Lasso	2D-MCP	LOG	NEG	1D-MCP	perm-MCP	genes
CD	rs12649928	4q34.3	181,646,381	2.39×10^{-03}	2.33×10^{-03}	0.034	-	-	9.14×10^{-08}	-	-	-	-	
CD	rs338852	5q22.2	111,865,496	7.83×10^{-04}	8.12×10^{-05}	0.976	3.06×10^{-06}	3.24×10^{-06}	2.68×10^{-08}	2.02×10^{-06}	-	-	-	EPB41L4A, FLJ11235, APC
CD	rs7820917	8p23.1	9,647,867	6.01×10^{-05}	3.84×10^{-06}	0.0176	-	-	-	-	4×10^{-07}	-	-	TNKS, LOC157627, MSRA
CD	rs17663476	8q21.13	83,251,023	4.99×10^{-05}	4.15×10^{-05}	0.99	5.19×10^{-06}	3.63×10^{-06}	1×10^{-08}	5.03×10^{-06}	-	1.59×10^{-05}	-	
CD	rs12548876	8q23.1	107,477,663	1.35×10^{-03}	2.22×10^{-04}	0.0547	-	-	2.87×10^{-08}	-	-	-	-	OXR1, ABRA
CD	rs9783122	10q25.1	106,766,397	3.98×10^{-04}	1.55×10^{-05}	0.973	1.14×10^{-05}	4.12×10^{-06}	6.47×10^{-11}	3.53×10^{-06}	1.82×10^{-08}	1.21×10^{-06}	-	SORCS3
CD	rs11031006	11p14.1	30,226,527	4.99×10^{-04}	1.77×10^{-05}	0.987	1.07×10^{-05}	1.55×10^{-05}	6.32×10^{-08}	7.16×10^{-06}	-	1.32×10^{-05}	-	KCNA4, FSHB, C11orf46, MPPED2
CD	rs17224453	14q32.12	91,928,907	2.04×10^{-03}	1.84×10^{-03}	0.0152	-	-	1.08×10^{-08}	-	-	-	-	C14orf159, GPR68, CCDC88C, SMEK1, C14orf184, CATSPERB
CD	rs4789523	17q25.3	76,012,853	2.61×10^{-04}	6.35×10^{-05}	0.936	6.2×10^{-05}	5.35×10^{-05}	5.24×10^{-09}	2.89×10^{-05}	-	-	-	11 genes
CD	rs4807569	19p13.3	1,123,377	6.06×10^{-06}	5.41×10^{-06}	0.975	8.46×10^{-07}	7.21×10^{-07}	3.67×10^{-09}	1.14×10^{-06}	8.63×10^{-06}	7.68×10^{-07}	-	29 genes
CD	rs1759092	19q13.11	34,676,537	1.06×10^{-03}	1.46×10^{-04}	0.894	-	-	9.31×10^{-09}	-	-	-	-	LSM14A, KIAA0355, GPI, PDCD2L, UBA2, WTIP
CD	rs3126001	Xq23	112,327,813	6.15×10^{-05}	6.15×10^{-05}	0.176	1.06×10^{-05}	4.79×10^{-06}	2.58×10^{-09}	4.77×10^{-05}	-	-	-	AMOT

Table A.7: Additional associations for rheumatoid arthritis identified by PMR methods but not a single marker analysis

				Method										
disease	SNP	chromosome	position	SMA	Conditional	VBAY	Lasso	Adaptive Lasso	2D-MCP	LOG	NEG	1D-MCP	perm-MCP	genes
RA	rs11162922	1p31.1	80,572,057	6.57×10^{-06}	6.45×10^{-06}	0.068	2.03×10^{-06}	1.36×10^{-05}	7.29×10^{-08}	1.56×10^{-06}	4.1×10^{-05}	1.65×10^{-06}	-	
RA	rs9872427	3p26.1	7,164,222	6.16×10^{-04}	6.16×10^{-04}	0.0104	1.33×10^{-05}	3.84×10^{-05}	7.99×10^{-08}	2.03×10^{-05}	-	-	-	GRM7
RA	rs702347	6q24.3	147,887,568	7.25×10^{-04}	1.07×10^{-04}	0.175	1.86×10^{-07}	1.44×10^{-07}	5.86×10^{-06}	1.58×10^{-06}	8.41×10^{-07}	1.58×10^{-06}	-	STXBP5, SAMD5
RA	rs11761231	7q32.3	131,370,038	4.55×10^{-06}	4.55×10^{-06}	0.173	1.19×10^{-05}	6.12×10^{-06}	3.13×10^{-08}	4.94×10^{-06}	1.9×10^{-05}	1.01×10^{-05}	-	MKLN1, PODXL
RA	rs4133002	8q13.3	72,718,580	1.24×10^{-04}	1.24×10^{-04}	-	-	-	9.46×10^{-09}	-	-	-	-	MSC, TRPA1
RA	rs2769190	9q33.1	120,942,581	4.89×10^{-03}	2.82×10^{-04}	-	-	-	1.89×10^{-08}	-	-	-	-	
RA	rs2514189	11q12.1	57,943,327	5.08×10^{-06}	5.08×10^{-06}	0.0837	1.85×10^{-07}	4.99×10^{-08}	1.89×10^{-12}	1.82×10^{-07}	2.44×10^{-09}	3.94×10^{-07}	-	12 genes
RA	rs11068744	12q24.23	118,364,177	5.86×10^{-04}	2.45×10^{-04}	0.0914	5.8×10^{-04}	1.92×10^{-03}	3.95×10^{-08}	6.75×10^{-05}	4.92×10^{-05}	1.24×10^{-04}	-	KSR2, RFC5, WSB2, VSIG10, PEBP1, TAOK3
RA	rs17365438	15q24.3	76,833,837	6.95×10^{-04}	1.53×10^{-05}	0.0717	5.17×10^{-07}	5.7×10^{-07}	5.96×10^{-10}	2.58×10^{-07}	3.4×10^{-07}	7.48×10^{-07}	-	ETFA, ISL2, SCAPER
RA	rs4889989	17q25.3	78,151,859	2.61×10^{-03}	8.45×10^{-04}	-	-	-	4.03×10^{-09}	-	-	-	-	10 genes
RA	rs6526752	Xp22.2	14,606,404	1.31×10^{-04}	3×10^{-05}	-	9.87×10^{-06}	5.24×10^{-06}	3.39×10^{-08}	3.47×10^{-06}	2.08×10^{-05}	-	-	GLRA2, FANCB, MOSPD2

Table A.8: Additional associations for type 1 diabetes identified by PMR methods but not a single marker analysis

				Method										
disease	SNP	chromosome	position	SMA	Conditional	VBAY	Lasso	Adaptive Lasso	2D-MCP	LOG	NEG	1D-MCP	perm-MCP	genes
T1D	rs577193	1q43	236,802,015	2.32×10^{-04}	2.32×10^{-04}	0.0483	2.9×10^{-04}	3.29×10^{-04}	6.09×10^{-08}	1.09×10^{-04}	-	-	-	EDARADD, LGALS8, HEATR1, ACTN2, MTR
T1D	rs10865679	3p13	73,781,216	1.96×10^{-02}	2.2×10^{-03}	-	-	-	5.75×10^{-08}	-	-	-	-	PDZRN3
T1D	rs4707786	6q13	71,058,226	5.11×10^{-04}	1.75×10^{-04}	0.375	1.14×10^{-05}	8.68×10^{-06}	3.32×10^{-08}	1.78×10^{-05}	-	3.5×10^{-05}	-	COL19A1, COL9A1, FAM135A, C6orf57
T1D	rs760500	6q27	170,725,386	5.24×10^{-04}	1.66×10^{-05}	0.133	1.55×10^{-07}	2.04×10^{-07}	6.63×10^{-10}	5.32×10^{-08}	-	1.77×10^{-07}	-	LOC154449, DLL1, FAM120B, PSMB1, TBP, PDCD2
T1D	rs346613	8q13.3	73,242,672	2.6×10^{-03}	2×10^{-05}	0.152	5.5×10^{-06}	3.22×10^{-06}	5.33×10^{-08}	2.32×10^{-06}	-	3.79×10^{-06}	-	TRPA1, KCNB2
T1D	rs10759987	9q33.1	121,364,133	9.98×10^{-05}	9.98×10^{-05}	0.0377	4.61×10^{-05}	5.62×10^{-05}	4.58×10^{-09}	5.07×10^{-05}	-	7.32×10^{-05}	-	
T1D	rs2666236	10p11.22	33,418,871	4.46×10^{-05}	2.24×10^{-06}	0.0374	1.12×10^{-04}	9.66×10^{-05}	3.88×10^{-08}	3.08×10^{-05}	-	1.04×10^{-04}	-	C10orf68, ITGB1, NRP1
T1D	rs4938390	11q23.3	117,358,312	9.24×10^{-05}	3.08×10^{-05}	0.0356	2.21×10^{-06}	7.65×10^{-07}	8.48×10^{-06}	4.78×10^{-07}	-	8.86×10^{-07}	-	7 genes
T1D	rs17756934	12q23.3	107,949,264	2.34×10^{-02}	1.65×10^{-04}	0.0127	3.27×10^{-06}	8.64×10^{-06}	7.99×10^{-12}	8×10^{-06}	-	3.76×10^{-06}	-	BTBD11, PWP1, PRDM4, ASCL4
T1D	rs341557	13q21.2	60,367,013	4.45×10^{-05}	4.45×10^{-05}	-	-	-	2.58×10^{-09}	2×10^{-05}	-	-	-	DIAPH3
T1D	rs1357809	18q12.1	27,656,198	1.8×10^{-02}	5.1×10^{-04}	0.011	-	-	6.04×10^{-08}	6.69×10^{-05}	-	-	-	

A.3 Effective degrees of freedom for the linear mixed model

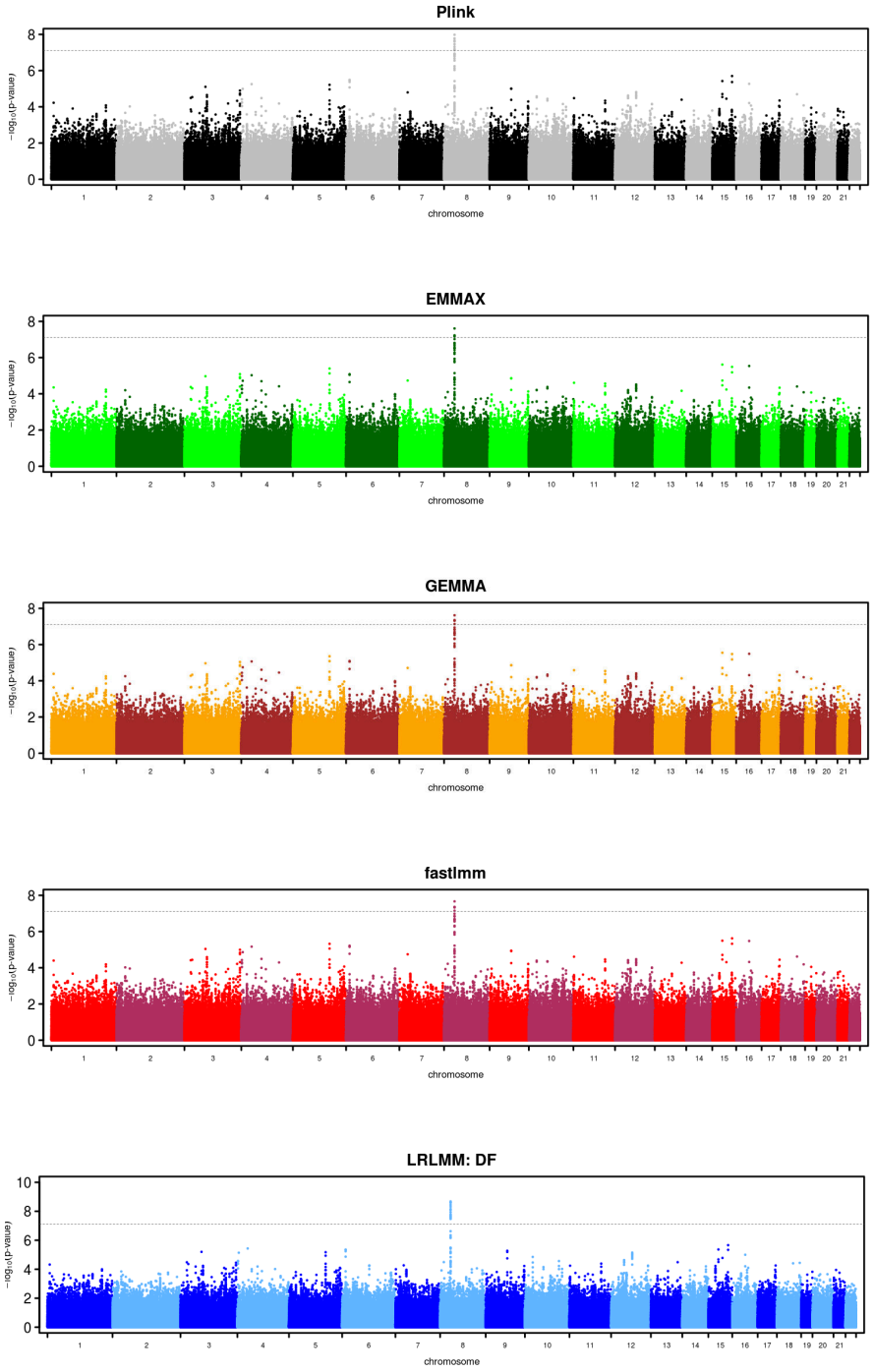
Show that $tr(\mathbf{H}) = \sum_i \frac{s_i^2}{s_i^2+\delta}$, letting $\delta = \frac{\sigma_e^2}{\sigma_a^2}$ and s_i^2 be the i^{th} eigen-value of $\mathbf{K}$. The derivation follows from standard properties of matrix algebra and orthonormal matrices in the eigen-decomposition.

$$\begin{aligned}
\mathbf{H} &= \mathbf{K}(\mathbf{K}+\mathbf{I}\delta)^{-1} \\
&= \mathbf{K}(\mathbf{U}\mathbf{S}^2\mathbf{U}^T + \mathbf{U}\mathbf{U}^T\delta)^{-1} \\
&= \mathbf{K}\left[\mathbf{U}(\mathbf{S}^2+\mathbf{I}\delta)\mathbf{U}^T\right]^{-1} \\
&= \mathbf{K}\mathbf{U}(\mathbf{S}^2+\mathbf{I}\delta)^{-1}\mathbf{U}^T \\
&= \mathbf{U}\mathbf{S}^2\mathbf{U}^T\mathbf{U}(\mathbf{S}^2+\mathbf{I}\delta)^{-1}\mathbf{U}^T \\
&= \mathbf{U}\mathbf{S}^2(\mathbf{S}^2+\mathbf{I}\delta)^{-1}\mathbf{U}^T
\end{aligned}$$

$$\begin{aligned}
tr(\mathbf{H}) &= tr[\mathbf{U}\mathbf{S}^2(\mathbf{S}^2+\mathbf{I}\delta)^{-1}\mathbf{U}^T] \\
&= tr[\mathbf{S}^2(\mathbf{S}^2+\mathbf{I}\delta)^{-1}\mathbf{U}^T\mathbf{U}] \\
&= tr[\mathbf{S}^2(\mathbf{S}^2+\mathbf{I}\delta)^{-1}] \\
&= \sum_i \frac{s_i^2}{s_i^2+\delta}
\end{aligned}$$

We note that evaluating df_e for the low rank linear mixed model has the same form since no assumption of positivity was made for the values of s_i^2.

Figure A.23: Manhattan plots for HDL cholesterol in Europeans from the Multi-Ethnic Study of Atherosclerosis (MESA) using Plink, EMMAX, GEMMA, fastlmm and our low rank linear mixed model sorting by degrees of freedom from fitting each principal component individually (LRLMM-DF).



BIBLIOGRAPHY

[1] Albert, J. H., Chib, S., Association, S., and Jun, N., 1993. Bayesian Analysis of Binary and Polychotomous Response Data. Journal of the American Statistical Association 88 (422), 669.

[2] Alexander, D. H. and Lange, K., 2011. Stability selection for genome-wide association. Genetic Epidemiology 35 (7), 722–8.

[3] Anderson, C. A., Boucher, G., Lees, C. W., Franke, A., D'Amato, M. et al., 2011. Meta-analysis identifies 29 additional ulcerative colitis risk loci, increasing the number of confirmed associations to 47. Nature Genetics 43 (3), 246–52.

[4] Anderson, C. A., Pettersson, F. H., Barrett, J. C., Zhuang, J. J., Ragoussis, J. et al., 2008. Evaluating the effects of imputation on the power, coverage, and cost efficiency of genome-wide SNP platforms. American Journal of Human Genetics 83 (1), 112–9.

[5] Antoniadis, A., Gijbels, I., and Nikolova, M., 2009. Penalized likelihood regression for generalized linear models with non-quadratic penalties. Annals of the Institute of Statistical Mathematics 63 (3), 585–615.

[6] Aschard, H., Chen, J., Cornelis, M., Chibnik, L., Karlson, E. et al., 2012. Inclusion of Gene-Gene and Gene-Environment Interactions Unlikely to Dramatically Improve Risk Prediction for Complex Diseases. American Journal of Human Genetics, 1–11.

[7] Asimit, J. and Zeggini, E., 2010. Rare variant association analysis methods for complex traits. Annual Review of Genetics 44, 293–308.

[8] Astle, W. and Balding, D. J., 2009. Population Structure and Cryptic Relatedness in Genetic Association Studies. Statistical Science 24 (4), 451–471.

[9] Ayers, K. L. and Cordell, H. J., 2010. SNP Selection in genome-wide and candidate gene studies via penalized logistic regression. Genetic Epidemiology 34 (8), 879–91.

[10] Barrett, J. C., Clayton, D. G., Concannon, P., Akolkar, B., Cooper, J. D. et al., 2009. Genome-wide association study and meta-analysis find that over 40 loci affect risk of type 1 diabetes. Nature Genetics 41 (6), 703–707.

[11] Barrett, J. C., Hansoul, S., Nicolae, D. L., Cho, J. H., Duerr, R. H. et al., 2008. Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease. Nature Genetics 40 (8), 955–62.

[12] Benson, D. a., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., and Wheeler, D. L., 2005. GenBank. Nucleic Acids Research 33 (Database issue), D34–8.

[13] Bild, D. E., 2002. Multi-Ethnic Study of Atherosclerosis: Objectives and Design. American Journal of Epidemiology 156 (9), 871–881.

[14] Boyd, S. S. and Vandenberghe, L., 2004. Convex Optimization.

[15] Breheny, P. and Huang, J., 2009. Penalized Methods for Bi-level variable selection. Statistics and Its Interface 2, 369–380.

[16] Breheny, P. and Huang, J., 2009. Penalized Methods for Bi-level variable selection. Statistics and Its Interface 2, 369–380.

[17] Breheny, P. and Huang, J., 2011. Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection. The Annals of Applied Statistics 5 (1), 232–253.

[18] Browning, S. R. and Browning, B. L., 2007. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. American Journal of Human Genetics 81 (5), 1084–97.

[19] Browning, S. R. and Browning, B. L., 2011. Population structure can inflate SNP-based heritability estimates. American Journal of Human Genetics 89 (1), 191–3.

[20] Buhlmann, P. and van der Geer, S., 2011. Statistics for high-dimensional data. Springer-Verlag, New York.

[21] Cantor, R. M., Lange, K., and Sinsheimer, J. S., 2010. Prioritizing GWAS results: A review of statistical methods and recommendations for their application. American Journal of Human Genetics 86 (1), 6–22.

[22] Carbonetto, P. and Stephens, M., 2011. Scalable variational inference for Bayesian variable selection in regression, and its accuracy in genetic association studies. Bayesian Analysis 6 (4), 1–42.

[23] Castelloe, J. M. and Brien, R. G. O., 2001. Power and Sample Size Determination for Linear Models A. SAS Users' Group International 26, 240.

[24] Cénit, M. C., Alcina, A., Márquez, A., Mendoza, J. L., Díaz-Rubio, M. et al., 2010. STAT3 locus in inflammatory bowel disease and multiple sclerosis susceptibility. Genes and Immunity 11 (3), 264–8.

[25] Chanock, S. J., Manolio, T., Boehnke, M., Boerwinkle, E., Hunter, D. J. et al., 2007. Replicating genotype-phenotype associations. Nature 447 (7145), 655–660.

[26] Chen, J. and Chen, Z., 2008. Extended Bayesian information criteria for model selection with large model spaces. Biometrika 95 (3), 759–771.

[27] Chen, L. S., Hutter, C. M., Potter, J. D., Liu, Y., Prentice, R. L. et al., 2010. Insights into colon cancer etiology via a regularized approach to gene set analysis of GWAS data. American Journal of Human Genetics 86 (6), 860–71.

[28] Cho, S., Kim, H., Oh, S., Kim, K., and Park, T., 2009. Elastic-net regularization approaches for genome-wide association studies of rheumatoid arthritis. BMC Proceedings 3 (Suppl 7), S25.

[29] Chung, A. S., Guan, Y.-j., Yuan, Z.-l., Albina, J. E., and Chin, Y. E., 2005. Ankyrin repeat and SOCS box 3 (ASB3) mediates ubiquitination and degradation of tumor necrosis factor receptor II. Molecular and Cellular Biology 25 (11), 4716–26.

[30] Cooper, J. D., Smyth, D. J., Smiles, A. M., Plagnol, V., Walker, N. M. et al., 2008. Meta-analysis of genome-wide association study data identifies additional type 1 diabetes risk loci. Nature Genetics 40 (12), 1399–1401.

[31] Coste, A., Dubuquoy, L., Barnouin, R., Annicotte, J.-S., Magnier, B. et al., 2007. LRH-1-mediated glucocorticoid synthesis in enterocytes protects against inflammatory bowel disease. Proceedings of the National Academy of Sciences of the United States of America 104 (32), 13098–103.

[32] Cupples, L. A., Arruda, H. T., Benjamin, E. J., D'Agostino, R. B., Demissie, S. et al., 2007. The Framingham Heart Study 100K SNP genome-wide association study resource: overview of 17 phenotype working group reports. BMC Medical Genetics 8 Suppl 1, S1.

[33] Daly, A. K., 2010. Genome-wide association studies in pharmacogenomics. Nature Reviews Genetics 11 (4), 241–6.

[34] Devlin, B. and Roeder, K., 1999. Genomic control for association studies. Biometrics 55 (4), 997–1004.

[35] Donoho, D. L. and Johnstone, J. M., 1994. Ideal spatial adaptation by wavelet shrinkage. Biometrika 81 (3), 425–455.

[36] Dubois, P. C. a., Trynka, G., Franke, L., Hunt, K. a., Romanos, J. et al., 2010. Multiple common variants for celiac disease influencing immune gene expression. Nature Genetics 42 (4), 295–302.

[37] Efron, B. and Hastie, T., 2004. Least angle regression. The Annals of statistics 32 (2), 407–499.

[38] Ehret, G. B., Munroe, P. B., Rice, K. M., Bochud, M., Johnson, A. D. et al., 2011. Genetic variants in novel pathways influence blood pressure and cardiovascular disease risk. Nature 478 (7367), 103–9.

[39] Eleftherohorinou, H., Hoggart, C. J., Wright, V. J., Levin, M., and Coin, L. J. M., 2011. Pathway-driven gene stability selection of two rheumatoid arthritis GWAS identifies and validates new susceptibility genes in receptor mediated signalling pathways. Human Molecular Genetics 20 (17), 3494–3506.

[40] Ellinghaus, D., Ellinghaus, E., Nair, R. P., Stuart, P. E., Esko, T. o. et al., 2012. Combined Analysis of Genome-wide Association Studies for Crohn Disease and Psoriasis Identifies Seven Shared Susceptibility Loci. American Journal of Human Genetics 90 (4), 636–647.

[41] Fan, J., 2010. A selective overview of variable selection in high dimensional feature space. Statistica Sinica 20, 101–148.

[42] Fan, J. and Li, R., 2001. Variable selection via nonconcave penalized likelihood and its oracle properties. Journal of the American Statistical Association 96, 1348–1360.

[43] Fan, J. and Lv, J., 2008. Sure independence screening for ultrahigh dimensional feature space. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 70 (5), 849–911.

[44] Fan, J. and Song, R., 2010. Sure independence screening in generalized linear models with NP-dimensionality. The Annals of Statistics 38 (6), 3567–3604.

[45] Franke, A., Hampe, J., Rosenstiel, P., Becker, C., Wagner, F. et al., 2007. Systematic association mapping identifies NELL1 as a novel IBD disease gene. PLoS ONE 2 (1), e691.

[46] Franke, A., McGovern, D. P. B., Barrett, J. C., Wang, K., Radford-Smith, G. L. et al., 2010. Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. Nature Genetics 42 (12), 1118–25.

[47] Friedman, J., Hastie, T., Höfling, H., and Tibshirani, R., 2007. Pathwise coordinate optimization. Annals of Applied Statistics 1 (2), 302–332.

[48] Friedman, J., Hastie, T., and Tibshirani, R., 2010. Regularization paths for generalized linear models via coordinate descent. Journal of Statistical Software 33 (1).

[49] Furberg, H., Kim, Y., Dackor, J., Boerwinkle, E., Franceschini, N. et al., 2010. Genome-wide meta-analyses identify multiple loci associated with smoking behavior. Nature Genetics 42 (5), 441–447.

[50] Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B., 2004. Bayesian Data Analysis. Chapman & Hall/CRC.

[51] Genkin, A., Lewis, D. D., and Madigan, D., 2007. Large-Scale Bayesian Logistic Regression for Text Categorization. Technometrics 49 (3), 291–304.

[52] Ghaoui, L. E., Viallon, V., and Rabbani, T., 2011. Safe Feature Elimination for the LASSO and Sparse Supervised Learning Problems. arXiv:1009.4219.

[53] Gregersen, P. K., Amos, C. I., Lee, A. T., Lu, Y., Remmers, E. F. et al., 2009. REL, encoding a member of the NF-kappaB family of transcription factors, is a newly defined risk locus for rheumatoid arthritis. Nature Genetics 41 (7), 820–3.

[54] Griffin, J. and Brown, P., 2005. Alternative prior distributions for variable selection with very many more variables than observations. Technical Report. University of Kent.

[55] Griffin, J. E. and Brown, P. J., 2007. Bayesian adaptive lassos with non-convex penalization. Technical Report. University of Warwick.

[56] Griffin, J. E. and Brown, P. J., 2011. Bayesian Hyper-Lassos With Non-Convex Penalization. Australian & New Zealand Journal of Statistics 53 (4), 423–442.

[57] Guan, Y. and Stephens, M., 2011. Bayesian variable selection regression for genome-wide association studies and other large-scale problems. The Annals of Applied Statistics 5 (3), 1780–1815.

[58] Hahn, L. W., Ritchie, M. D., and Moore, J. H., 2003. Multifactor dimensionality reduction software for detecting gene-gene and gene-environment interactions. Bioinformatics 19 (3), 376–382.

[59] Hakonarson, H., Grant, S. F. a., Bradfield, J. P., Marchand, L., Kim, C. E. et al., 2007. A genome-wide association study identifies KIAA0350 as a type 1 diabetes gene. Nature 448 (7153), 591–4.

[60] Hakonarson, H., Qu, H.-q., Bradfield, J. P., Marchand, L., Kim, C. E. et al., 2008. A novel susceptibility locus for type 1 diabetes on Chr12q13 identified by a genome-wide association study. Diabetes 57 (4), 1143–6.

[61] Halls, M. L., van der Westhuizen, E. T., Bathgate, R. A. D., and Summers, R. J., 2007. Relaxin family peptide receptors–former orphans reunite with their parent ligands to activate multiple signalling pathways. British Journal of Pharmacology 150 (6), 677–91.

[62] Hamosh, A., Scott, A. F., Amberger, J. S., Bocchini, C. a., and McKusick, V. a., 2005. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. Nucleic Acids Research 33 (Database issue), D514–7.

[63] Hastie, T., Taylor, J., Tibshirani, R., and Walther, G., 2007. Forward stagewise regression and the monotone lasso. Electronic Journal of Statistics 1, 1–29.

[64] Hastie, T., Tibshirani, R., and Friedman, J., 2009. The Elements of Statistical Learning, 2nd Edition. Springer Series in Statistics. Springer, New York, NY.

[65] He, Q. and Lin, D.-Y., 2011. A variable selection method for genome-wide association studies. Bioinformatics (Oxford, England) 27 (1), 1–8.

[66] Henderson, C., 1984. Applications of Linear Models in Animal Breeding. University of Guelph, Guelph, Ontario.

[67] Hesterberg, T., Choi, N. H., Meier, L., and Fraley, C., 2008. Least angle and L1 penalized regression: A review. Statistics Surveys 2, 61–93.

[68] Hindorff, L. A., Sethupathy, P., Junkins, H. A., Ramos, E. M., Mehta, J. P. et al., 2009. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proceedings of the National Academy of Sciences of the United States of America 106 (23), 9362–7.

[69] Hirschhorn, J. N. and Gajdos, Z. K. Z., 2011. Genome-wide association studies: results from the first few years and potential implications for clinical medicine. Annual Review of Medicine 62, 11–24.

[70] Hoerl, A. E. and Kennard, R. W., 1970. Ridge Regression : Biased Problems Nonorthogonal Estimation for Nonorthogonal Problems. Technometrics 12 (1), 55–67.

[71] Hoeting, J., Madigan, D., and Raftery, A., 1999. Bayesian model averaging: A tutorial. Statistical Science 14 (4), 382–401.

[72] Hoggart, C. J., Whittaker, J. C., De Iorio, M., and Balding, D. J., 2008. Simultaneous analysis of all SNPs in genome-wide and re-sequencing association studies. PLoS Genetics 4 (7), e1000130.

[73] Huang, H., Chanda, P., Alonso, A., Bader, J. S., and Arking, D. E., 2011. Gene-based tests of association. PLoS Genetics 7 (7), e1002177.

[74] Huber, P. J., 1973. Robust Regression: Asymptotics, Conjectures and Monte Carlo. The Annals of Statistics 1 (5), 799–821.

[75] Hunter, D. R., Lange, K., Unter, D. R. H., and Ange, K. L., 2004. A Tutorial on MM Algorithms. The American Statistician 58 (1), 30–37.

[76] Ising, M., Lucae, S., Binder, E. B., Bettecken, T., Uhr, M. et al., 2009. A genomewide association study points to multiple loci that predict antidepressant drug treatment outcome in depression. Archives of General Psychiatry 66 (9), 966–75.

[77] Jaakkola, T. and Jordan, M., 2000. Bayesian parameter estimation via variational methods. Statistics and Computing 10, 25–37.

[78] Jablonski, K. A., McAteer, J. B., de Bakker, P. I. W., Franks, P. W., Pollin, T. I. et al., 2010. Common variants in 40 genes assessed for diabetes incidence and response to metformin and lifestyle intervention in the diabetes prevention program. Diabetes 59 (10), 2672–81.

[79] Janss, L., de Los Campos, G., Sheehan, N., and Sorensen, D., 2012. Inferences from genomic models in stratified populations. Genetics 192 (2), 693–704.

[80] Jostins, L., Ripke, S., Weersma, R. K., Duerr, R. H., McGovern, D. P. et al.,

2012. Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. Nature 491 (7422), 119–24.

[81] Julià, A., Ballina, J., Cañete, J. D., Balsa, A., Tornero-Molina, J. et al., 2008. Genome-wide association study of rheumatoid arthritis in the Spanish population: KLF12 as a risk locus for rheumatoid arthritis susceptibility. Arthritis and Rheumatism 58 (8), 2275–86.

[82] Kanehisa, M., Goto, S., Sato, Y., Furumichi, M., and Tanabe, M., 2012. KEGG for integration and interpretation of large-scale molecular data sets. Nucleic Acids Research 40 (Database issue), D109–14.

[83] Kang, H. M., Sul, J. H., Service, S. K., Zaitlen, N. A., Kong, S.-y. et al., 2010. Variance component model to account for sample structure in genome-wide association studies. Nature Genetics 42 (4), 348–54.

[84] Kang, H. M., Zaitlen, N. A., Wade, C. M., Kirby, A., Heckerman, D. et al., 2008. Efficient control of population structure in model organism association mapping. Genetics 178 (3), 1709–23.

[85] Keinan, A. and Clark, A. G., 2012. Recent Explosive Human Population Growth Has Resulted in an Excess of Rare Genetic Variants. Science 336 (6082), 740–743.

[86] Kelleher, S., McCormick, N., Velasquez, V., and Lopez, V., 2011. Zinc in Specialized Secretory Tissues: Roles in the Pancreas, Prostate, and Mammary Gland. Advances in Nutrition: An International Review Journal 2 (2), 101.

[87] Kenny, E. E., Kim, M., Gusev, A., Lowe, J. K., Salit, J. et al., 2011. Increased power of mixed models facilitates association mapping of 10 loci

for metabolic traits in an isolated population. Human Molecular Genetics 20 (4), 827–39.

[88] Khor, B., Gardet, A., and Xavier, R. J., 2011. Genetics and pathogenesis of inflammatory bowel disease. Nature 474 (7351), 307–17.

[89] Kim, S. H. S., Cleary, M. M., Fox, H. S. H., Chantry, D., and Sarvetnick, N., 2002. CCR4-bearing T cells participate in autoimmune diabetes. Journal of Clinical Investigation 110 (11), 1675–1686.

[90] Kobberup, S., Nyeng, P., Juhl, K., Hutton, J., and Jensen, J., 2007. ETS-family genes in pancreatic development. Developmental Dynamics 236 (11), 3100–10.

[91] Kooperberg, C. and Ruczinski, I., 2005. Identifying interacting SNPs using Monte Carlo logic regression. Genetic Epidemiology 28, 157–170.

[92] Korte, A., Vilhjálmsson, B. J., Segura, V., Platt, A., Long, Q. et al., 2012. A mixed-model approach for genome-wide association studies of correlated traits in structured populations. Nature Genetics 44 (9), 1066–71.

[93] Kraja, A. T., Vaidya, D., Pankow, J. S., Goodarzi, M. O., Assimes, T. L. et al., 2011. A bivariate genome-wide approach to metabolic syndrome: STAMPEED consortium. Diabetes 60 (4), 1329–39.

[94] Kutner, M. H., Neter, J., Nachtsheim, C. J., and Li, W., 2004. Applied Linear Statistical Models, 5th Edition. McGraw-Hill.

[95] Lango Allen, H., Estrada, K., Lettre, G., Berndt, S. I., Weedon, M. N. et al., 2010. Hundreds of variants clustered in genomic loci and biological pathways affect human height. Nature 467 (7317), 832–838.

[96] Lawson, D. J. and Falush, D., 2012. Population identification using genetic data. Annual Review of Genomics and Human Genetics 13, 337–61.

[97] Lawson, D. J., Hellenthal, G., Myers, S., and Falush, D., 2012. Inference of population structure using dense haplotype data. PLoS Genetics 8 (1), e1002453.

[98] Lee, S., Emond, M. J., Bamshad, M. J., Barnes, K. C., Rieder, M. J. et al., 2012. Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. American Journal of Human Genetics 91 (2), 224–37.

[99] Lee, S., Wright, F. A., and Zou, F., 2011. Control of population stratification by correlation-selected principal components. Biometrics 67 (3), 967–74.

[100] Lee, Y., Nelder, J. A., and Pawitan, Y., 2006. Generalized Linear Models with Random Effects: Unified Analysis via H-likelihood. Chapman & Hall/CRC.

[101] Lehne, B., Lewis, C. M., and Schlitt, T., 2011. From SNPs to genes: disease association at the gene level. PloS ONE 6 (6), e20133.

[102] Li, B. and Leal, S., 2008. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. American Journal of Human Genetics 83, 311–321.

[103] Li, J., Das, K., Fu, G., Li, R., and Wu, R., 2010. The Bayesian Lasso for Genome-wide Association Studies. Bioinformatics 27 (4), 516–523.

[104] Li, J., Horstman, B., and Chen, Y., 2011. Detecting epistatic effects in association studies at a genomic level based on an ensemble approach. Bioinformatics 27 (13), i222–9.

[105] Li, M.-X., Kwan, J. S. H., and Sham, P. C., 2012. HYST: a hybrid set-based test for genome-wide association studies, with application to protein-protein interaction-based association analysis. American Journal of Human Genetics 91 (3), 478–88.

[106] Libioulle, C., Louis, E., Hansoul, S., Sandor, C., Farnir, F. et al., 2007. Novel Crohn disease locus identified by genome-wide association maps to a gene desert on 5p13.1 and modulates expression of PTGER4. PLoS Genetics 3 (4), e58.

[107] Lichten, L. a. and Cousins, R. J., 2009. Mammalian zinc transporters: nutritional and physiologic regulation. Annual Review of Nutrition 29, 153–76.

[108] Lippert, C., Listgarten, J., Liu, Y., Kadie, C. M., Davidson, R. I. et al., 2011. FaST linear mixed models for genome-wide association studies. Nature Methods 8 (10), 833–5.

[109] Listgarten, J., Lippert, C., Kadie, C. M., Davidson, R. I., Eskin, E. et al., 2012. Improved linear mixed models for genome-wide association studies. Nature Methods 9 (6), 525–6.

[110] Liu, H., 2003. The role of apoptosis in rheumatoid arthritis. Current Opinion in Pharmacology 3 (3), 317–322.

[111] Liu, J. Z., Mcrae, A. F., Nyholt, D. R., Medland, S. E., Wray, N. R. et al., 2010. A Versatile Gene-Based Test for Genome-wide Association Studies. American Journal of Human Genetics 87 (1), 139–145.

[112] Lockhart, R., Taylor, J., Tibshirani, R., and Tibshirani, R., 2013. A significance test for the lasso. arXive:1301.7161.

[113] Logsdon, B. A., Carty, C. L., Reiner, A. P., Dai, J. Y., and Kooperberg, C., 2012. A novel variational Bayes multiple locus Z-statistic for genome-wide association studies with Bayesian model averaging. Bioinformatics 28 (13), 1738–44.

[114] Logsdon, B. A., Hoffman, G. E., and Mezey, J. G., 2010. A variational Bayes algorithm for fast and accurate multiple locus genome-wide association analysis. BMC Bioinformatics 11, 58.

[115] Logsdon, B. A., Hoffman, G. E., and Mezey, J. G., 2012. Mouse obesity network reconstruction with a variational Bayes algorithm to employ aggressive false positive control. BMC Bioinformatics 13 (1), 53.

[116] Lou, X.-Y., Chen, G.-B., Yan, L., Ma, J. Z., Zhu, J. et al., 2007. A generalized combinatorial approach for detecting gene-by-gene and gene-by-environment interactions with application to nicotine dependence. American Journal of Human Genetics 80 (6), 1125–37.

[117] Lynch, M. and Walsh, B., 1998. Genetics and Analysis of Quantitative Traits. Sinauer Associates.

[118] Ma, H., Schadt, E. E., Kaplan, L. M., and Zhao, H., 2011. COSINE: COndition-SpecIfic sub-NEtwork identification using a global optimization method. Bioinformatics 27 (9), 1290–8.

[119] Ma, L., Brautbar, A., Boerwinkle, E., Sing, C. F., Clark, A. G. et al., 2012. Knowledge-Driven Analysis Identifies a Gene-Gene Interaction Affecting High-Density Lipoprotein Cholesterol Levels in Multi-Ethnic Populations. PLoS Genetics 8 (5), e1002714.

[120] Madsen, B. E. and Browning, S. R., 2009. A groupwise association test for rare mutations using a weighted sum statistic. PLoS Genetics 5 (2), e1000384.

[121] Maher, B., 2008. Personal genomes: The case of the missing heritability. Nature 456 (7218), 18–21.

[122] Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A. et al., 2009. Finding the missing heritability of complex diseases. Nature 461 (7265), 747–53.

[123] Mathieson, I. and McVean, G., 2012. Differential confounding of rare and common variants in spatially structured populations. Nature Genetics 44 (3), 243–6.

[124] Mazumder, R., Friedman, J., and Hastie, T., 2011. SparseNet : Coordinate Descent With Nonconvex Penalties. Journal of the American Statistical Association (5), 1–14.

[125] McCullagh, P. and Nelder, J. A., 1989. Generalized Linear Models, 2nd Edition. Chapman & Hall / CRC, New York.

[126] McGovern, D. P. B., Jones, M. R., Taylor, K. D., Marciante, K., Yan, X. et al., 2010. Fucosyltransferase 2 (FUT2) non-secretor status is associated with Crohn's disease. Human Molecular Genetics 19 (17), 3468–3476.

[127] McKinney, B. a., Crowe, J. E., Guo, J., and Tian, D., 2009. Capturing the spectrum of interaction effects in genetic association studies by simulated evaporative cooling network analysis. PLoS Genetics 5 (3), e1000432.

[128] McVean, G., 2009. A genealogical interpretation of principal components analysis. PLoS Genetics 5 (10), e1000686.

[129] Meier, L. and Geer, S. V. D., 2008. The group lasso for logistic regression. Journal of the Royal Statistical Society. Series B (Methodological) 70, 53–71.

[130] Meinshausen, N. and Bühlmann, P., 2010. Stability selection. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 72 (4), 417–473.

[131] Meinshausen, N., Meier, L., and Buhlmann, P., 2009. p -Values for High-Dimensional Regression. Journal of the American Statistical Association 104 (488), 1671–1681.

[132] Mitin, N., Rossman, K. L., and Der, C. J., 2005. Signaling interplay in Ras superfamily function. Current Biology 15 (14), R563–74.

[133] Moore, J. H., Asselbergs, F. W., and Williams, S. M., 2010. Bioinformatics challenges for genome-wide association studies. Bioinformatics 26 (4), 445–55.

[134] Morris, A. P. and Zeggini, E., 2010. An evaluation of statistical approaches to rare variant analysis in genetic association studies. Genetic Epidemiology 34 (2), 188–93.

[135] Motsinger-Reif, A. A., Jorgenson, E., Relling, M. V., Kroetz, D. L., Weinshilboum, R. et al., 2010. Genome-wide association studies in pharmacogenomics: successes and lessons. Pharmacogenetics and genomics.

[136] Nelson, M. R., Wegmann, D., Ehm, M. G., Kessner, D., St Jean, P. et al., 2012. An abundance of rare functional variants in 202 drug target genes sequenced in 14,002 people. Science 337 (6090), 100–4.

[137] Noti, M., Sidler, D., and Brunner, T., 2009. Extra-adrenal glucocorticoid synthesis in the intestinal epithelium: more than a drop in the ocean? Seminars in Immunopathology 31 (2), 237–48.

[138] O'Brien, S. M. and Dunson, D. B., 2004. Bayesian multivariate logistic regression. Biometrics 60 (3), 739–746.

[139] Oldham, K., Myland, J., and Spanier, J., 2008. An Atlas of Functions: with Equator, the Atlas Function Calculator. Springer Verlag, New York.

[140] Orozco, G., McAllister, K., and Eyre, S., 2011. Genetics of rheumatoid arthritis: GWAS and beyond. Open Access Rheumatology: Research and Reviews 3, 31–46.

[141] Park, J.-H., Wacholder, S., Gail, M. H., Peters, U., Jacobs, K. B. et al., 2010. Estimation of effect size distribution from genome-wide association studies and implications for future discoveries. Nature Genetics 42 (7), 570–5.

[142] Parkes, M., Barrett, J. C., Prescott, N. J., Tremelling, M., Anderson, C. a. et al., 2007. Sequence variants in the autophagy gene IRGM and multiple other replicating loci contribute to Crohn's disease susceptibility. Nature Genetics 39 (7), 830–2.

[143] Patel, M. N., Halling-Brown, M. D., Tym, J. E., Workman, P., and Al-Lazikani, B., 2013. Objective assessment of cancer genes for drug discovery. Nature Reviews Drug Discovery 12 (1), 35–50.

[144] Patterson, N., Price, A. L., and Reich, D., 2006. Population structure and eigenanalysis. PLoS Genetics 2 (12), e190.

[145] Pinheiro, J. and Bates, D., 2000. Mixed-Effects Models in S and S-PLUS. Springer, New York.

[146] Pirinen, M., Donnelly, P., and Spencer, C., 2012. Efficient computation with a linear mixed model on large-scale data sets with applications to genetic studies. Annals of Applied Statistics In press.

[147] Plenge, R. M., Seielstad, M., Padyukov, L., Lee, A. T., Remmers, E. F. et al., 2007. TRAF1-C5 as a risk locus for rheumatoid arthritis–a genomewide study. The New England Journal of Medicine 357 (12), 1199–209.

[148] Portnoy, S., 1984. Asymptotic Behavior of M-Estimators of p Regression Parameters when pˆ2/n is Large. I. Consistency. The Annals of Statistics 12 (4), 1298–1309.

[149] Prabhu, S. and Pe'er, I., 2012. Ultrafast genome-wide scan for SNP-SNP interactions in common complex disease. Genome Research 22 (11), 2230–40.

[150] Price, A., Patterson, N., Plenge, R., Weinblatt, M., Shadick, N. et al., 2006. Principal components analysis corrects for stratification in genome-wide association studies. Nature Genetics 38 (8), 904–909.

[151] Price, A. L., Kryukov, G. V., de Bakker, P. I. W., Purcell, S. M., Staples, J. et al., 2010. Pooled association tests for rare variants in exon-resequencing studies. American Journal of Human Genetics 86 (6), 832–8.

[152] Price, A. L., Zaitlen, N. A., Reich, D., and Patterson, N., 2010. New approaches to population stratification in genome-wide association studies. Nature Reviews Genetics 11 (7), 459–463.

[153] Punta, M., Coggill, P. C., Eberhardt, R. Y., Mistry, J., Tate, J. et al., 2012. The Pfam protein families database. Nucleic Acids Research 40 (Database issue), D290–301.

[154] Purcell, S., Neale, B., Toddbrown, K., Thomas, L., Ferreira, M. et al., 2007. PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. American Journal of Human Genetics 81 (3), 559–575.

[155] Raelson, J. V., Little, R. D., Ruether, A., Fournier, H., Paquin, B. et al., 2007. Genome-wide association study for Crohn's disease in the Quebec Founder Population identifies multiple validated disease loci. Proceedings of the National Academy of Sciences of the United States of America 104 (37), 14747–14752.

[156] Rakitsch, B., Lippert, C., Stegle, O., and Borgwardt, K., 2013. A Lasso multi-marker mixed model for association mapping with population structure correction. Bioinformatics 29 (2), 206–14.

[157] Rasmussen, C. and Williams, C., 2006. Gaussian Processes in Machine Learning. MIT Press, Cambridge, Massachusetts.

[158] Rioux, J. D., Xavier, R. J., Taylor, K. D., Silverberg, M. S., Goyette, P. et al., 2007. Genome-wide association study identifies new susceptibility loci for Crohn disease and implicates autophagy in disease pathogenesis. Nature Genetics 39 (5), 596–604.

[159] Ritchie, M. D., 2011. Using biological knowledge to uncover the mystery in the search for epistasis in genome-wide association studies. Annals of Human Genetics 75 (1), 172–82.

[160] Ritchie, M. D., Hahn, L. W., and Moore, J. H., 2003. Power of multifactor dimensionality reduction for detecting gene-gene interactions in the presence of genotyping error, missing data, phenocopy, and genetic heterogeneity. Genetic Epidemiology 24 (2), 150–7.

[161] Romanoski, C. E., Lee, S., Kim, M. J., Ingram-drake, L., Plaisier, C. L. et al., 2010. Systems Genetics Analysis of Gene-by-Environment Interactions in Human Cells. American Journal of Human Genetics 86 (3), 399–410.

[162] Ruppert, D., Wand, M., and Carroll, R., 2003. Semiparametric Regression. Cambridge University Press, New York.

[163] Sanseau, P., Agarwal, P., Barnes, M. R., Pastinen, T., Richards, J. B. et al., 2012. Use of genome-wide association studies for drug repositioning. Nature biotechnology 30 (4), 317–20.

[164] Sawcer, S., Hellenthal, G., Pirinen, M., Spencer, C. C. a., Patsopoulos, N. a. et al., 2011. Genetic risk and a primary role for cell-mediated immune mechanisms in multiple sclerosis. Nature 476 (7359), 214–219.

[165] Schaid, D. J., 2010. Genomic Similarity and Kernel Methods II: Methods for Genomic Information. Human Heredity 70 (2), 132–140.

[166] Schifano, E. D., Strawderman, R. L., and Wells, M. T., 2010. Majorization-Minimization algorithms for nonsmoothly penalized objective functions. Electronic Journal of Statistics 4, 1258–1299.

[167] Schüpbach, T., Xenarios, I., Bergmann, S., and Kapur, K., 2010. FastEpistasis: A high performance computing solution for quantitative trait epistasis. Bioinformatics 26 (11), 1468–1469.

[168] Schwarz, G., 1978. Estimating the Dimension of a Model. The Annals of Statistics 6 (2), 461–464.

[169] Searle, S., Casella, G., and McCulloch, C., 1992. Variance Components. John Wiley & Sons, USA.

[170] Segura, V., Vilhjálmsson, B. J., Platt, A., Korte, A., Seren, U. et al., 2012. An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations. Nature Genetics 44 (7), 825–830.

[171] Sevilla, A., Santos, C. R., Barcia, R., Vega, F. M., and Lazo, P. A., 2004. c-Jun phosphorylation by the human vaccinia-related kinase 1 (VRK1) and its cooperation with the N-terminal kinase of c-Jun (JNK). Oncogene 23 (55), 8950–8.

[172] Sorensen, D. and Gianola, D., 2002. Likelihood, Bayesian and MCMC Methods in Quantitative Genetics. Springer, New York.

[173] Speed, D., Hemani, G., Johnson, M. R., and Balding, D. J., 2012. Improved Heritability Estimation from Genome-wide SNPs. American Journal of Human Genetics 91 (6), 1011–1021.

[174] Speliotes, E. K., Willer, C. J., Berndt, S. I., Monda, K. L., Thorleifsson, G. et al., 2010. Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. Nature Genetics 42 (11), 937–948.

[175] Spencer, C. A., 2009. Designing genome-wide association studies: Sample size, power imputation, and the choice of genotyping chip. PLoS Genetics 5e1000477.

[176] Stahl, E. A., Raychaudhuri, S., Remmers, E. F., Xie, G., Eyre, S. et al., 2010. Genome-wide association study meta-analysis identifies seven new rheumatoid arthritis risk loci. Nature Genetics 42 (6), 508–14.

[177] Stelzer, G., Dalah, I., Stein, T. I., Satanower, Y., Rosen, N. et al., 2011. In-silico human genomics with GeneCards. Human Genomics 5 (6), 709–17.

[178] Stranger, B. E., Stahl, E. a., and Raj, T., 2011. Progress and promise of genome-wide association studies for human complex trait genetics. Genetics 187 (2), 367–83.

[179] Suhre, K. and Gieger, C., 2012. Genetic variation in metabolic phenotypes: study designs and applications. Nature Reviews Genetics 13 (11), 759–69.

[180] Suhre, K., Shin, S.-Y., Petersen, A.-K., Mohney, R. P., Meredith, D. et al., 2011. Human metabolic individuality in biomedical and pharmaceutical research. Nature 477 (7362), 54–60.

[181] Suzuki, K., Kumanogoh, A., and Kikutani, H., 2008. Semaphorins and their receptors in immune cell interactions. Nature Immunology 9 (1), 17–23.

[182] Svishcheva, G. R., Axenovich, T. I., Belonogova, N. M., van Duijn, C. M., and Aulchenko, Y. S., 2012. Rapid variance components-based method for whole-genome association analysis. Nature Genetics 44 (10), 1166–70.

[183] Takeuchi, F., McGinnis, R., Bourgeois, S., Barnes, C., Eriksson, N. et al., 2009. A genome-wide association study confirms VKORC1, CYP2C9, and CYP4F2 as principal genetic determinants of warfarin dose. PLoS Genetics 5 (3), e1000433.

[184] Tennessen, J. A., Bigham, A. W., O'Connor, T. D., Fu, W., Kenny, E. E. et al., 2012. Evolution and functional impact of rare coding variation from deep sequencing of human exomes. Science 337 (6090), 64–9.

[185] Teslovich, T. M., Musunuru, K., Smith, A. V., Edmondson, A. C., Stylianou, I. M. et al., 2010. Biological, clinical and population relevance of 95 loci for blood lipids. Nature 466 (7307), 707–713.

[186] Thomas, D., 2010. Gene-environment-wide association studies: emerging approaches. Nature Reviews Genetics 11 (4), 259–272.

[187] Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society. Series B (Methodological) 58 (1), 267–288.

[188] Tibshirani, R., Bien, J., Friedman, J., Hastie, T., Simon, N. et al., 2012. Strong rules for discarding predictors in lasso-type problems. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 74 (2), 245–266.

[189] Tibshirani, R. J., 2012. The Lasso Problem and Uniqueness. arXive:1206.0313.

[190] Todd, J. a., Walker, N. M., Cooper, J. D., Smyth, D. J., Downes, K. et al., 2007. Robust associations of four new chromosome regions from genome-wide analyses of type 1 diabetes. Nature Genetics 39 (7), 857–64.

[191] van Steen, K., 2012. Travelling the world of gene-gene interactions. Briefings in Bioinformatics 13 (1), 1–19.

[192] Vignal, C. M., Bansal, A. T., and Balding, D. J., 2011. Using Penalised Logistic Regression to Fine Map HLA Variants for Rheumatoid Arthritis. Annals of Human Genetics 75 (6), 655–64.

[193] Visscher, P., Brown, M., McCarthy, M., and Yang, J., 2012. Five Years of GWAS Discovery. American Journal of Human Genetics 90 (1), 7–24.

[194] Voight, B. F., Peloso, G. M., Orho-Melander, M., Frikke-Schmidt, R., Barbalic, M. et al., 2012. Plasma HDL cholesterol and risk of myocardial infarction: a mendelian randomisation study. Lancet 380 (9841), 572–80.

[195] Voight, B. F., Scott, L. J., Steinthorsdottir, V., Morris, A. D. P., Dina, C. et al., 2010. Twelve type 2 diabetes susceptibility loci identified through large-scale association analysis. Nature Genetics 42 (7), 579–589.

[196] Wainwright, M. J. and Jordan, M. I., 2007. Graphical Models, Exponential Families, and Variational Inference. Foundations and Trends in Machine Learning 1 (12), 1–305.

[197] Wainwright, M. J., Ravikumar, P., and Lafferty, J. D., 2007. High-Dimensional Graphical Model Selection Using l˜ 1-Regularized Logistic Regression. Advances in Neural Information Processing Systems 19, 1465.

[198] Wang, H., 2009. Forward Regression for Ultra-High Dimensional Variable Screening. Journal of the American Statistical Association 104 (488), 1512–1524.

[199] Wang, H., Li, R., and Tsai, C.-L., 2007. Tuning parameter selectors for the smoothly clipped absolute deviation method. Biometrika 94 (3), 553–568.

[200] Wang, K., Li, M., and Bucan, M., 2007. Pathway-based approaches for analysis of genomewide association studies. American Journal of Human Genetics 81 (6), 1278–1283.

[201] Wang, K., Li, M., and Hakonarson, H., 2010. Analysing biological pathways in genome-wide association studies. Nature Reviews Genetics 11 (12), 843–854.

[202] Wasserman, L. and Roeder, K., 2009. High-dimensional variable selection. The Annals of Statistics 37 (5A), 2178–2201.

[203] Waterworth, D. M., Ricketts, S. L., Song, K., Chen, L., Zhao, J. H. et al., 2010. Genetic variants influencing circulating lipid levels and risk of coro-

nary artery disease. Arteriosclerosis, Thrombosis, and Vascular Biology 30 (11), 2264–76.

[204] Weissglas-Volkov, D. and Pajukanta, P., 2010. Genetic causes of high and low serum HDL-cholesterol. Journal of Lipid Research 51 (8), 2032–57.

[205] Wellcome Trust Case Control Consortium, 2007. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 447 (7145), 661–78.

[206] Wood, S., 2006. Generalized Additive Models: An Introduction with R. Chapman & Hall/CRC, Boca Raton, Florida.

[207] Wu, C., DeWan, A., Hoh, J., and Wang, Z., 2011. A comparison of association methods correcting for population stratification in case-control studies. Annals of Human Genetics 75 (3), 418–27.

[208] Wu, M. C., Lee, S., Cai, T., Li, Y., Boehnke, M. et al., 2011. Rare-Variant Association Testing for Sequencing Data with the Sequence Kernel Association Test. American Journal of Human Genetics 89 (1), 82–93.

[209] Wu, T. T., Chen, Y. F., Hastie, T., Sobel, E., and Lange, K., 2009. Genome-wide association analysis by lasso penalized logistic regression. Bioinformatics 25 (6), 714–21.

[210] Wu, T. T. and Lange, K., 2008. Coordinate descent algorithms for lasso penalized regression. Annals of Applied Statistics 2 (1), 224–244.

[211] Wu, Y., Boos, D. D., and Stefanski, L. a., 2007. Controlling Variable Selection by the Addition of Pseudovariables. Journal of the American Statistical Association 102 (477), 235–243.

[212] Wullaert, A., Heyninck, K., Janssens, S., and Beyaert, R., 2006. Ubiquitin: tool and target for intracellular NF-kappaB inhibitors. Trends in Immunology 27 (11), 533–40.

[213] Yang, A., Kaghad, M., Caput, D., and McKeon, F., 2002. On the shoulders of giants: p63, p73 and the rise of p53. Trends in Genetics 18 (2), 90–95.

[214] Yang, C., Wan, X., Yang, Q., Xue, H., and Yu, W., 2010. Identifying main effects and epistatic interactions from large-scale SNP data via adaptive group Lasso. BMC Bioinformatics 11 Suppl 1, S18.

[215] Yang, J., Benyamin, B., McEvoy, B. P., Gordon, S., Henders, A. K. et al., 2010. Common SNPs explain a large proportion of the heritability for human height. Nature Genetics 42 (7), 565–9.

[216] Yang, J., Lee, S. H., Goddard, M. E., and Visscher, P. M., 2011. GCTA: a tool for genome-wide complex trait analysis. American Journal of Human Genetics 88 (1), 76–82.

[217] Yang, J., Manolio, T. a., Pasquale, L. R., Boerwinkle, E., Caporaso, N. et al., 2011. Genome partitioning of genetic variation for complex traits using common SNPs. Nature Genetics 43 (6), 519–25.

[218] Yu, W., Gwinn, M., Clyne, M., Yesupriya, A., and Khoury, M. J., 2008. A navigator for human genome epidemiology. Nature Genetics 40 (2), 124–5.

[219] Yuan, M. and Lin, Y., 2006. Model selection and estimation in regression with grouped variables. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 68 (1), 49–67.

[220] Zawistowski, M., Gopalakrishnan, S., Ding, J., Li, Y., Grimm, S. et al., 2010. Extending rare-variant testing strategies: analysis of noncoding sequence and imputed genotypes. American Journal of Human Genetics 87 (5), 604–17.

[221] Zhang, C.-H., 2010. Nearly unbiased variable selection under minimax concave penalty. The Annals of Statistics 38 (2), 894–942.

[222] Zhang, M., Zhang, D., and Wells, M. T., 2008. Variable selection for large p small n regression models with incomplete data: mapping QTL with epistases. BMC Bioinformatics 9, 251.

[223] Zhang, X., Huang, S., Zou, F., and Wang, W., 2010. TEAM: efficient two-locus epistasis tests in human genome-wide association study. Bioinformatics 26 (12), i217–27.

[224] Zhang, Y., Li, R., Tsai, C.-L., Hang, Y. Z., Runze, L. I. et al., 2010. Regularization Parameter Selections via Generalized Information Criterion. Journal of the American Statistical Association 105 (489), 312–323.

[225] Zhang, Z., Ersoz, E., Lai, C.-Q., Todhunter, R. J., Tiwari, H. K. et al., 2010. Mixed linear model approach adapted for genome-wide association studies. Nature Genetics 42 (4), 355–60.

[226] Zhao, P. and Yu, B., 2006. On model selection consistency of Lasso. The Journal of Machine Learning Research 7, 2541–2563.

[227] Zhou, H., Sehl, M. E., Sinsheimer, J. S., and Lange, K., 2010. Association Screening of Common and Rare Genetic Variants by Penalized Regression. Bioinformatics 26 (19), 2375–2382.

[228] Zhou, K., Bellenguez, C., Spencer, C. C. a., Bennett, A. J., Coleman, R. L. et al., 2011. Common variants near ATM are associated with glycemic response to metformin in type 2 diabetes. Nature Genetics 43 (2), 117–20.

[229] Zhou, K. and Pearson, E. R., 2013. Insights from genome-wide association studies of drug response. Annual Review of Pharmacology and Toxicology 53, 299–310.

[230] Zhou, X. and Stephens, M., 2012. Genome-wide efficient mixed-model analysis for association studies. Nature Genetics 44 (7), 821–4.

[231] Zou, F., Huang, H., Lee, S., and Hoeschele, I., 2010. Nonparametric Bayesian variable selection with applications to multiple quantitative trait loci mapping with epistasis and gene-environment interaction. Genetics 186 (1), 385–94.

[232] Zou, H., 2006. The adaptive lasso and its oracle properties. Journal of the American Statistical Association 101 (476), 1418–1429.